Evolutionary and Paleobiochemical Analyses of Heme Peroxidases

by

Noeleen B. Loughran B. Sc. (Hons) Biotechnology

A thesis presented to Dublin City University for the
Degree of Doctor of Philosophy

Supervisors:
Dr. Mary J. O'Connell, Dr. Brendan O'Connor, Dr. Ciarán Ó Fagáin
School of Biotechnology
Dublin City University

September 2010

Declaration

'I hereby declare that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.'

Signed: _____

I.D. Number: 52541301

Date: 22 September 2010

Changing

"It is not so much what man is that counts as it is what he ventures to make of himself. To make the leap he must do more than disclose himself; he must risk a certain amount of confusion. Then, as soon as he does catch a glimpse of a different kind of life, he needs to find some way of overcoming the paralyzing moment of threat, for this is the instant when he wonders who he really is - whether he is what he just was or is what he is about to be."

George Kelly

For my mother

Acknowledgements

To my supervisory committee, Dr. Mary O'Connell, Dr. Ciarán Ó Fagáin and Dr. Brendan O'Connor, a heartfelt thank you for giving me the opportunity to carry out this research, for your constant support, guidance and enthusiasm over the last few years. It has been a privilege to work for and with you!

I would like to thank Prof. William Nauseef and the Iowa Inflammation Program for providing me with the opportunity to work in your laboratories and for sharing your expertise. A special thanks to the School of Biotechnology and the Benson family for making possible this collaboration.

The BME lab, Mary, Tom, Claire and Mark, a huge thanks for all your help and encouragement and of course all the laughs – good times! In a nutshell :) keep her lit!

To all in the protein lab, thanks for all your insightful suggestions and all the fun times.

Thanks to all my colleagues and friends in Biotechnology and Chemistry for making my postgrad experience so enjoyable.

Many thanks to the girls, Sinéad, Eva, Susan and Aileen. You have been an unbelievable help and support to me and have made the last four years the best.

A special thanks to Yvonne for being a truly great friend.

To my girls, Anna, Emma, Fiona and Vicki, thanks for being so understanding and supportive and for always being there. You have been and always will be fantastic friends!

Last, but by no means least, I would like to sincerely thank my family, Daddy, Sinéad, Irene, Bernard and Garry. You have shown constant support and encouragement and have been an inspiration. Words can't thank you enough!

Abbreviations

AA	Amino Acids
ABTS	2,2'-azino-bis(3-ethyl-benzthiazoline-6-sulphonic acit)
AD	Alzheimer's Diseaese
AIC	Akaike Information Criterion
Ala	Alanine (A)
APF	3'-(*p*-aminophenyl) fluorescin
Arg	Arginine (R)
Asn	Asparagine (N)
Asp	Aspartic Acid (D)
ATCC	American Type Culture Collection
β - ME	Beta- Mercaptoethanol
BCA	Bicinchoninic Acid
BCIP/NBT	5-Bromo-4-chloro-3-indolyl phosphate/Nitro Blue Tetrazolium
BEB	Bayes Empirical Bayes
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
Br^-	Bromide ion
BSA	Bovine Serum Albumin
C^{14}	Carbon 14
CO_2	Carbon Dioxide
CIP	Calf Intestinal Phosphatase
Cl^-	Chloride ion
CL – ELISA	Chemiluminescent Enzyme-linked Immunosorbent Assay
CLN	Calnexin
CRT	Calreticulin
Cys	Cysteine (C)
D	Data
DDC	Duplication-Degeneration-Complementation

dH_2O	Distilled Water
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethylsulphoxide
DNA	Deoxyribonucleic Acid
D_n	Non-synonymous substitutions per non-synonymous site
D_s	Synonymous substitutions per synonymous site
D_n/D_s	Rate of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site
E	Elution Fraction
EB/AO	Ethidium Bromide/Acridine Orange
EF-TU	Elongation Factor
EPO	Eosinophil Peroxidase
ER	Endoplasmic Reticulum
FT	Flow Through
G	Guanine
GP	Archetypal gene sequence (Grandparent)
Glu	Glutanic Acid (E)
Gln	Glutamine (Q)
H_2O	Water
H_2O_2	Hydrogen Peroxide
H-bond	Hydrogen Bond
HEK	Human Embryonal Kidney cells 293
His	Histidine (H)
HIV	Human Immunodeficiency Virus (Type 1 (I) and 2)
hLRTs	Hierarchical Likelihood Ratio Tests
HOCL	Hypochlorous Acid
HRP	Horseradish Peroxidase (isoenzymes C and A2)
I^-	Iodide ion
IMAC	Immobilized Metal Affinity Chromatography

IPTG	Isopropyl β-D-1-thiogalactopyranoside
JTT	Jones Taylor Thornton (substitution matrix)
k	Rate Constant
L	Likelihood Measurement
LB	Luria Bertani
LBA	Long Branch Attraction
Leu	Leucine (L)
LDL	Low Density Lipoprotein
LPO	Lactoperoxidase
LRTs	Likelihood Ratio Tests
Lys	Lysine (K)
M	Model
ME	Minimum Evolution
Met	Methionine (M)
MHP	Mammalian Heme Peroxidase
ML	Maximum Likelihood
MPO	Myeloperoxidase
mRNA	Messenger RNA
MRCA	Most Recent Common Ancestor
MSA	Multiple Sequence Alignment
MY	Million Years
MYO	Million Year Old
n	Number of Iterations
N_e	Effective Population Size
NEB	Naïve Empirical Bayes
NNI	Nearest Neighbour Interchange
NJ	Neighbour Joining
P	Probability
p	Level of Significance (p-value)

PAGE	Polyacrylamide Gel Electrophoresis
PAML	Phylogenetic Analysis by Maximum Likelihood
PAUP	Phylogenetic Analysis Using Parsimony
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
Phe	Phenylalanine (F)
PMNs	Polymorphonuclear Neutrophils
PMSF	Phenylmethylsulfonyl Fluoride
PPs	Posterior Probabilities
Pro	Proline (P)
PXDN	Peroxidasin
RMSD	Root Means Squared Deviation
RNA	Ribonucleic Acid
RZ	Reinheitzahl Number
S^{35}	Sulphur 35
Ser	Serine (S)
SBP	Soybean Peroxidase
SDS-PAGE	Sodium Dodecylsulphate Polyacrylamide Gel Electrophoresis
SEM	Standard Error Mean
sH_2O	Sterile Water
SIV	Simian Immunodeficiency Virus
SOC	Super Optimal Broth with Catabolite Repression
τ	tree
T	Thymine
$t_{1/2}$	Apparent Half Life
TBS	Tris Buffered Saline
Thr	Treonine (T)
TIOD	Total Iodide Organification Defect
TMB	Tetramethylbenzidine Dihydrochloride
TPO	Thyroid Peroxidase
TRIM	Tripartite Motif

tRNA	Transfer RNA
Trp	Tryptophan (W)
UPGMA	Unweighted Pair Group Method with Arithmetic
v	branch length
v/v	Volume/Volume
WT	Wild Type
w/v	Weight/Volume
ω	Rate of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site
Y	Tyrosine (Tyr)
θ	substitution model (evolutionary process)

List of Figures

List of Tables

Abstract

The focus of this thesis is to study the evolution of enzyme specificity, and the application of evolutionary theory to the design of enzymes with desirable characteristics for industry. The approach presented here, applied to the heme peroxidases, marries bioinformatic methods with evolutionary theory and biochemical validations.

Heme peroxidases catalyse the oxidation of a variety of electron donors by hydrogen peroxide. These enzymes can be classified into two major families that arose from independent evolutionary events; the plant and the animal peroxidases. The first results chapter, Chapter 2, deals with the animal (mammalian) heme peroxidases known collectively as the MHP. Four main superfamilies of MHP have been classified; myeloperoxidase (MPO), eosinophil peroxidase (EPO), lactoperoxidase (LPO) and thyroid peroxidase (TPO).These comprise a functionally diverse multigene family of enzymes associated with such diseases as asthma, Alzheimer's disease and inflammatory vascular disease. This study has determined how the enzymes in the multigene family of MHP are related. The order of gene duplication events has been traced, with an MPO-EPO-LPO most recent common ancestor (MRCA) arising from a gene duplication with extant TPO. A further duplication event gave rise to (i) the MPO-EPO MRCA, and (ii) the lineage leading to extant LPO. The final and most recent duplication of the MPO-EPO MRCA resulted in the extant MPO and EPO clades. This phylogeny was subsequently used to predict the amino acids that have most likely contributed to each of the diverse functions of MHP. Positively selected sites have been identified, through the use of Bayesian estimation, unique to all four MHP. Using MPO as a case study, *in vitro* analyses on the impact of mutating these positions, specifically mutants Y500F and L504T, indicates a disruption to the biosynthesis and loss of enzymatic activity in our mutants supporting our *in silico* predictions. This work is described in results Chapter 3. Finally, Chapter 4 details the analysis of ancestral protein reconstruction within the plant peroxidase gene family. The phylogeny of plant peroxidases had previously been resolved; this allowed for the generation of the ancestral enzyme, estimated age approx. 113 million years old. This enzyme was cloned, expressed and found to be active. Catalytic and stability properties of this unique enzyme have been ascertained. Together, these analyses provide a valuable insight into enzyme function through molecular evoultionary analyses of sequence data and serve to bridge the gap between protein sequence, structure, and function.

Chapter 1

Introduction

1.1 Heme Peroxidases

Heme peroxidases (EC 1.11.1.X) are a ubiquitous subset of enzymes capable of catalysing the oxidation of a variety of electron donors by hydrogen peroxide (H_2O_2). The basic enzymatic action of heme peroxidases follows the reaction:

$$\text{donor} + H_2O_2 = \text{oxidised donor} + 2H_2O \qquad Eqn. \quad 1$$

where the donor is an H_2O_2 oxidoreductase (Barman 1969). The classic peroxidase cycle, undertaken by all heme peroxidases, follows a series of oxidative reactions. The native ferric enzyme (porphyrin Fe(III)) form is oxidised to compound I (Fe(IV)=O and porphyrin radical cation), which in turn is reduced (one electron reduction) to form compound II (Fe(IV)=O), which, upon further reduction (one electron reduction) is converted back to the native state. Common reducing substrates include phenols and anilines. Both compound I and II are oxidised intermediate forms of the native enzyme and are powerful oxidants. In the presence of excess H_2O_2, the native enzyme state and compound II may be reduced to the resting state, compound III (Dunford 1999). An alternative to the peroxidase cycle is the halogenation cycle (Furtmüller et al. 2006). Compound I has the ability to oxidise halide ions including Cl^-, Br^- and I^-, resulting in the halogenating agents HOX, where X corresponds to Cl, Br or I. See Figure 1.1 for a schematic representation of the classic peroxidase and halogenation cycles.

Based on sequence homology peroxidases can be classified into two major families, the animal peroxidases and those peroxidases present in bacteria, fungi and plants, referred to hereafter as the plant peroxidase superfamily (Passardi et al. 2007a). Both animal and plant superfamilies are thought to have arisen from two independent evolutionary events (O'Brien 2000).

The conserved active site present in all heme peroxidases contains a heme-based prosthetic group. This prosthetic group in plant peroxidases is a ferriprotoporphyrin IX

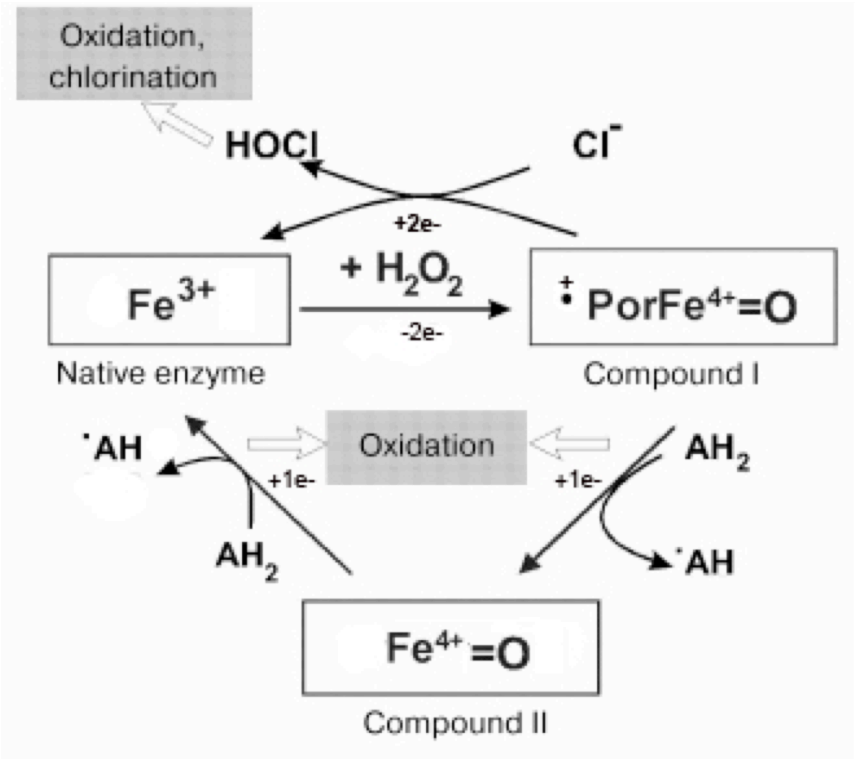


Figure 1.1. Peroxidation and halogenation cycles. The classic peroxidase cycle can be seen with the oxidation of native enzyme by hydrogen peroxide (H_2O_2) to form compound I, followed by further oxidation to compound II with reversion back to the native enzyme. Depending on substrate availability, after the oxidation of native enzyme to compound I, the halogenation pathway/cycle may be entered as depicted by the oxidation of compound I by the halide ion, Cl^-, resulting in the generation of hypochlorous acid (HOCl). This is the predominant pathway for MPO. AH_2 and $^{\bullet}AH$ represent the oxidisable substrate and its oxidized product respectively. Por = porphyrin, e- = electron. Adapted from (Arnhold 2004).

group whereas in animals it is a covalently bound heme; this difference in prosthetic groups is used to distinguish between plant and mammalian heme peroxidases (MHP) (Metcalfe et al. 2004). For MHP, the heme is joined *via* two linkages (Glu and Asp; conserved in all MHP) (Furtmüller et al. 2006). For one of the MHP members, myeloperoxidase (MPO), in addition to the two ester linkages, the heme is also joined via one sulfonium ion linkage at position 409 (Met) (Furtmüller et al. 2006). This unique threefold linkage is thought to be associated with MPO's unique chlorination activity (Kooter et al. 1999). Calcium ions are also linked to the heme-binding region via a network of hydrogen bonds, which are fundamental to the structure and in turn function of heme peroxidases (Dunford 1999; Furtmüller et al. 2006). The activity of peroxidases is dependent on the presence and correct conformation of the heme group (Neves-Petersen et al. 2007). See Figure 1.2 for illustration of the chemical structure of plant and animal heme groups and their different oxidative states.

1.1.1 Animal/Mammalian heme peroxidases

Four main superfamilies of mammalian heme peroxidases (MHP) have been classified; myeloperoxidase (MPO), eosinophil peroxidase (EPO), lactoperoxidase (LPO) and thyroid peroxidase (TPO). MPO, EPO and LPO have key roles in anti-microbial and innate immune responses, whereas, TPO is involved in thyroid hormone biosynthesis (Ruf, Carayon 2006). A study on the structure-function relationships of human heme peroxidases suggests that TPO evolved independently of MPO, EPO and LPO, and that these three members of the MHP share a common ancestral gene (Sakamaki et al. 2000; Sakamaki, Ueda, Nagata 2002; Furtmüller et al. 2006). The evolutionary relationships between these functionally diverse mammalian peroxidases can be resolved using molecular/sequence data. This analysis has been performed as outlined in results chapter 2.

Human MPO, EPO and LPO are all located within the same chromosomal region, the long arm (q) of chromosome 17, whereas human TPO is located on the short arm (p) of

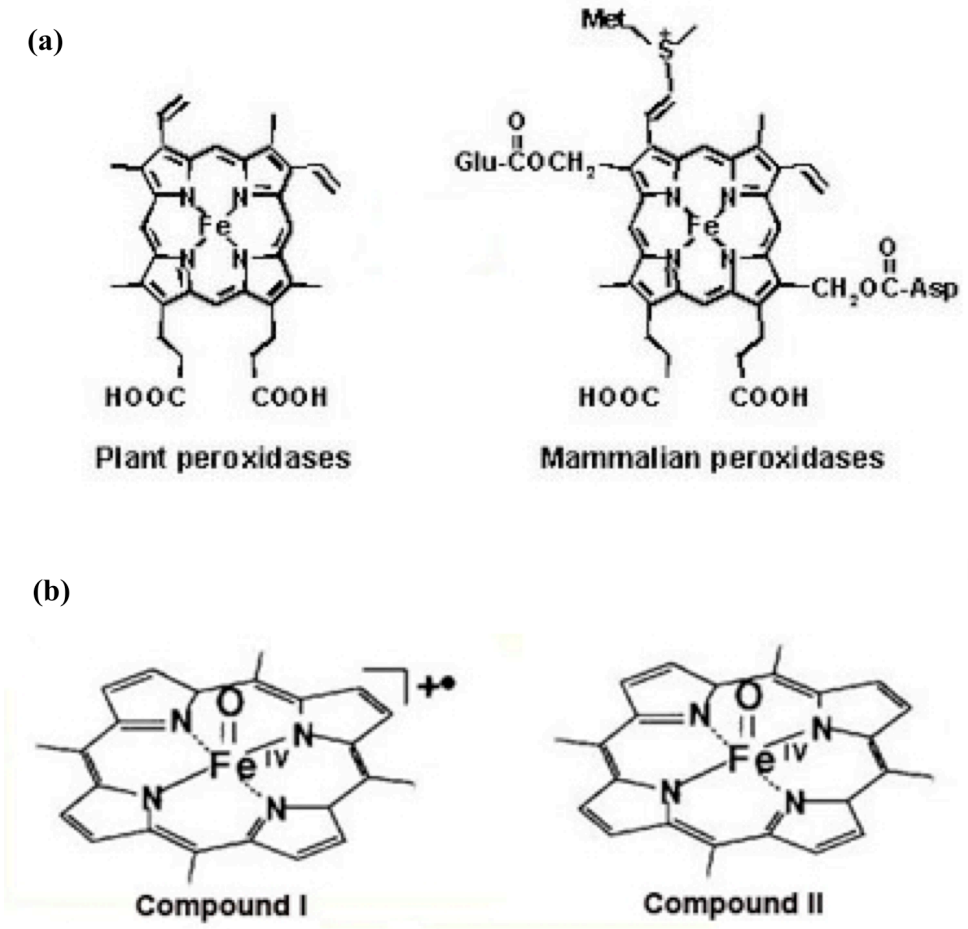


Figure 1.2. Peroxidase heme groups. (a) On the left is the prosthetic heme group present in plant heme peroxidases. On the right is the covalently linked heme group of animal/mammalian heme peroxidases with the conserved two ester linkages (Glu and Asp) shown and the sulfonium ion linkage present only in MPO (Met). (b) The oxidised heme group for the compound I and compound II intermediates can be seen on the left and on the right respectively. Adapted from **http://www2.le.ac.uk/departments/chemistry/people/academic-staff/prof-ravens-research-interests**.

chromosome 2. MHP proteins are expressed in various tissues, with their respective biological function dependent on cellular location (see Table1.1). Their chromosomal locations support the hypothesis that MPO, EPO and LPO share a most recent common ancestor (MRCA) and that this MRCA arose from a gene duplication event with the ancestor of extant TPO (Sakamaki et al. 2000; Sakamaki, Ueda, Nagata 2002; Furtmüller et al. 2006). This hypothesized relationship is depicted graphically in Figure 1.3.

1.1.2 *In vivo* biological function of mammalian heme peroxidases

TPO's primary function is in the biosynthesis of thyroid hormones, where it oxidises the naturally occurring iodide ion (I^-) allowing for the subsequent incorporation of iodine into thyroglobulin. This protein is then used by the thyroid gland to produce thyroid hormones (Ruf, Carayon 2006). The oxidants generated by MPO, EPO and LPO are crucial in the innate immune responses of the body, i.e. defending the body against disease. The roles of MPO, EPO and LPO in host defence have been extensively reviewed (Reiter 1978; deWit, vanHooydonk 1996; Hampton, Kettle, Winterbourn 1998; Meeusen, Balic 2000; Davies et al. 2008).

MPO is a major constituent of the azurophilic granules of polymorphonuclear neutrophils (PMNs). The majority of H_2O_2 generated by neutrophils is believed to be consumed during the peroxidation/halogenation action of MPO. The killing of bacteria by neutrophil-derived MPO is crucial in first-line defence against bacterial infection. Studies on MPO-knockout mice have shown increased bacterial infection susceptibility with respect to wild type mice (Aratani et al. 1999; Chen, Row, Hong 2002). In comparison to the other members of the MHP family, MPO has a unique function in oxidising chloride. This results in the generation of the potent cytotoxic agent hypochlorous acid (HOCl). Such chlorinating activity is considered to be MPO's predominant function in neutrophil-derived bactericidal action (Davies et al. 2008).

Table 1.1: Mammalian heme peroxidase features and functions. (Adapted from Clark 2000 and O'Brien 2000).

Superfamily (EC no.)	Chromosomal Location (Human)	Tissue Expression	Biological Function
MPO (1.11.1.7)	17	Neutrophils, mono-nuclear phagocytes	Microbicidal activity
EPO (1.11.1.7)	17	Eosinophils	Microbicidal activity
LPO (1.11.1.7)	17	Milk, saliva, tears and other secretions	Bacteriostatic and bactericidal activity
TPO (1.11.1.8)	2	Thyroid cell surface and cytoplasm	Thyroid hormone biosynthesis

MPO = Myeloperoxidase; EPO = Eosinophil peroxidase; LPO = Lactoperoxidase; TPO = Thyroid peroxidase.

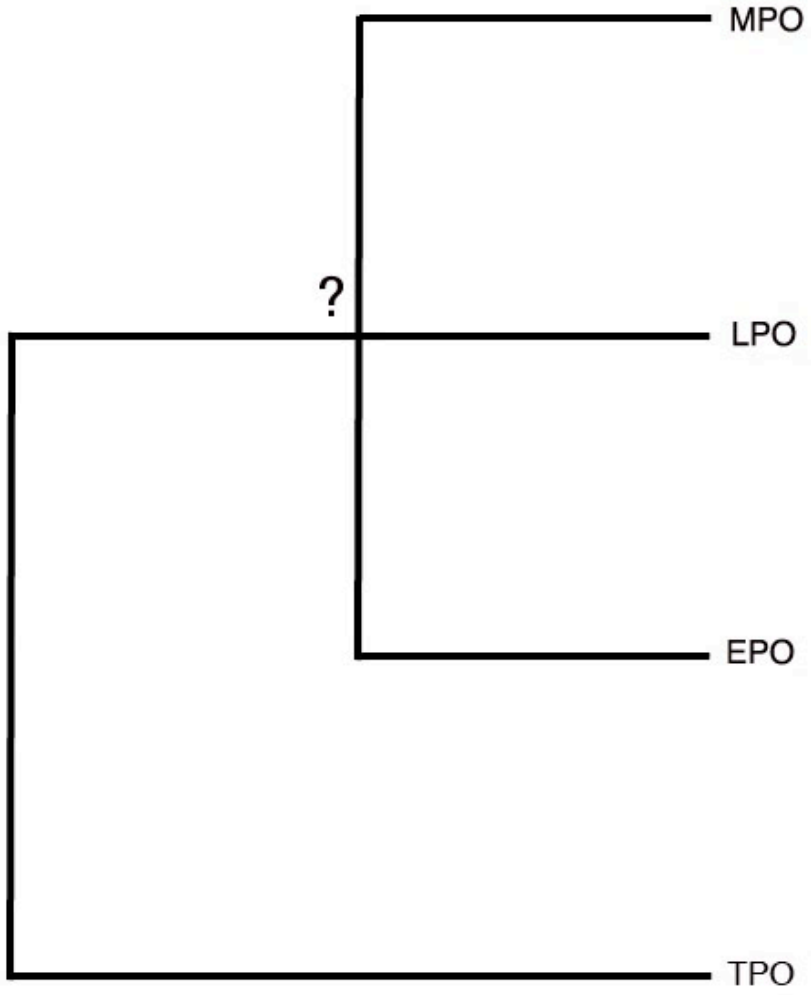


Figure 1.3. Hypothetical relationship between MHP. It has been hypothesised that that MPO, EPO and LPO share a most recent common ancestor (MRCA) and that this MRCA arose from a gene duplication event with the ancestor of extant TPO (Furtmüller *et al.* 2006). The questions remaining, which are addressed in this thesis (chapter 2), are (i) how are the three enzymes that are believed to share a MRCA relate to each other, (ii) what is the true evolutionary relationship of the entire MHP multigene family and (iii) how have their function diversified?

The white blood cells known as eosinophils contain EPO, and are known to play a major role in combating parasite and viral infections (Klebanoff, Coombs 1996; Meeusen, Balic 2000). EPO is exclusively expressed in eosinophils; see Table 1.1. Isolated EPO in the presence of H_2O_2 and halide ions has the ability to kill invading parasites and has also been shown to be virucidal towards human immunodeficiency virus type I (HIV-I) (Jong, Mahmoud, Klebanoff 1981; Klebanoff, Coombs 1996; Davies et al. 2008).

LPO is present in secretions from mammary, salivary and mucosal glands (Tenovuo 1985). LPO is known to play a major role in innate immune defence through the generation of natural bactericidal agents in milk, saliva and other secretions (Reiter 1978; Thomas et al. 1983; deWit, vanHooydonk 1996). LPO is also present in the airway mucosa and studies have highlighted its role in protecting the body against inhaled toxins and particles (Gerson et al. 2000; Conner, Salathe, Forteza 2002; Davies et al. 2008).

1.1.3 Health implications

Genetic abnormalities of MHP genes have been shown to result in deficient disease states. This is reflected by inherited MPO deficiency in the case of MPO and total iodide organification defect (TIOD) in the case of TPO. The genotypes of a large number of patients diagnosed with MPO deficiency have revealed the causative missense mutations (Nauseef, Brigham, Cogley 1994; Nauseef, Cogley, McCormick 1996; Romano et al. 1997; DeLeo et al. 1998; Ohashi et al. 2004; Persad et al. 2006; Goedken et al. 2007). In depth analyses on the cellular fate of these mutants *in vitro* has been carried out previously (Nauseef, Cogley, McCormick 1996; DeLeo et al. 1998; Goedken et al. 2007). Further details on MPO deficiency and associated polymorphisms are discussed in Chapter 3 of this thesis. Similarily, TPO deficiency is strongly linked to the heritable TIOD, which causes goitrous congenital hypothyroidism (Nascimento et al. 2003; Tajima, Tsubaki, Fujieda 2005). It has been highlighted that extensive analyses of the TPO gene is required for those diagnosed with TIOD to better understand this genetic defect (Fugazzola et al. 2005).

Although MPO, EPO and LPO have an important influence on the body's innate immune responses, their generated oxidants can also have an adverse effect by causing tissue damage (sometimes severe), resulting in the initiation and progression of various degenerative and inflammatory disease states such as multiple sclerosis and rheumatoid arthritis, as discussed by Petrides and Nauseef (2000) and Davies *et al.* (2008).

MPO is implicated in many pathologies including cardiovascular disease, Alzheimer's disease (AD) and some cancers (Reynolds et al. 1997; Reynolds et al. 2000; Lau, Baldus 2006). MPO activity has been detected in all grades of atherosclerotic lesions of humans, and has been used as a marker in predicting early risk myocardial infarctions (Daugherty et al. 1994; Heinecke 1999; Baldus et al. 2003; Brennan et al. 2003; Nicholls, Hazen 2005; Lau, Baldus 2006). Experimental analysis has detected high levels of MPO in the shoulder regions of the body, which is a common site for rupture (Daugherty et al. 1994). MPO-derived oxidants have been to shown to promote oxidation of low-density lipoprotein (LDL), which can enhance atherosclerosis (Hazell, Stocker 1993; Heinecke 1997). Studies have also indicated that MPO-derived oxidants (namely HOCl) can result in plaque erosion/rupture and thrombogenesis, as evidenced by the colocalisation of MPO and related oxidised proteins in the lesions of patients that have experienced sudden cardiac death (Sugiyama et al. 2001; Sugiyama et al. 2004).

High levels of MPO have been detected in the brains of patients with diseases such as AD and Parkinson's disease, suggesting that MPO-derived oxidants may be involved in the progression of neurodegenerative diseases (Reynolds et al. 1999; Green et al. 2004; Yap, Whiteman, Cheung 2007). Insoluble β-amyloid plaques deposits are characteristic of AD. MPO protein has been shown to colocalise with β-amyloid proteins, suggesting that MPO promotes the aggregation of these senile plaques and that its derived oxidants are potentially damaging to brain tissue (Reynolds et al. 1999; Green et al. 2004).

Sites of inflammation, which are clearly associated with cell killing by oxidants, may display a compensatory increase in division of the surviving cells and thus possess an increased risk in the development of cancer. MPO expression has been shown to act as a

prognostic marker for different myeloid leukemias (Matsuo et al. 2003). LPO has also been proposed to be involved in the development of cancers, specifically breast carcinogenesis, due to its associated oxidation of estrogenic hormones and of proteins present in breast milk (Josephy 1996; Cavalieri et al. 1997).

EPO and EPO-derived oxidants hold strong associations with asthma and allergic diseases, as do eosinophil cells themselves (Mitra, Slungaard, Hazen 2000). Levels of EPO are used as a clinical marker for determining the severity of asthma in patients (Rao et al. 1996; Sanz et al. 1997; Parra et al. 1999). Severe asthmatics have been shown to have significantly higher levels of EPO than those detected in mild to moderate patients, suggesting that EPO-derived oxidants may play an important role in lung damage. MPO has also been associated with severe asthmatic patients suffering from bacterial infections (Tauber et al. 1999).

In summary, although the potent cytotoxic oxidants generated by these enzymes are capable of bactericidal and virucidal action they have also been associated with many pathologies due to their oxidative damage to tissue. This adverse action has allowed for the use of their expression levels as clinical markers for the initiation and progression of many of their associated diseases.

1.1.4 PeroxiBase

PeroxiBase (http://peroxibase.isb-sib.ch) is a publicly available database that houses all known/available peroxidase (EC 1.11.1.X) sequences (complete and/or partial) and was the first repository devoted to a superfamily composed of multigene families (Bakalovic et al. 2006; Passardi et al. 2007b; Koua et al. 2009; Oliva et al. 2009). It was first established in 2004 at the University of Geneva, Switzerland as a repository for only class III plant peroxidases. The database centralized most of the annotated and non-annotated class III peroxidases, making them publicly available to the research community and providing putative functional information. Early versions of the database allowed external persons to deposit peroxidase sequences easily and directly. However,

this allowed for increased redundancy in the database, which prompted sequence entry restrictions in later versions, whereby PeroxiBase curators verify external contributions prior to inclusion in the repository.

Since its inception, PeroxiBase has been expanded with the aim of including all peroxidases, both heme (haem) and non-heme (non-haem), from prokaryotes and eukaryotes. At present, the database houses 7,118 peroxidases, and is updated periodically. PeroxiBase has allowed for the development of a consistent and standard nomenclature for the various peroxidases. Any discrepancies regarding the nomenclature in previous literature and databases have been accounted for by conserving previous identifiers within PeroxiBase. The database has seen a number of developments over the last five years that have provided useful tools and facilities for analysing the stored sequences. Researchers can browse the repository using specified criteria based on 'Classes', 'Organisms', 'Cellular localizations', 'Inducers', 'Repressors' and 'Tissue Types'. Sequences can be retrieved from the database in FASTA format for further independent analyses. Selected sequences may be searched against the database under the multi-criteria search tools and using the incorporated scanning tools (e.g. the standard sequence similarity search algorthim BLAST and PeroxiScan, which searches for peroxidase domains, are both fully integrated into the database). The database is linked to various other protein and DNA repositories (e.g. SWISS-PROT, NCBI) allowing for cross-referencing of data and routine updating/collating of sequences. PeroxiBase provides a powerful tool for efficiently analysing these ubiquitous enzymes.

1.1.5 Plant heme peroxidases

The other heme peroxidase superfamily, i.e. those of bacteria, fungi and plants, has been subdivided into three major groups: classes I (*peroxidases of prokaryotic origin*); II (*secreted fungal peroxidases*); and III (*secretory plant peroxidases*). Studies have been undertaken to determine the phylogenetic relationships within the plant heme peroxidases (Duroux, Welinder 2003; Zamocky 2004). Durox and Welinder sought to elucidate the evolutionary relationship amongst the class III secretory plant peroxidase,

of which horseradish peroxidase (HRP) and soybean peroxidases (SBP) are members (Duroux, Welinder 2003). The relationships between the class III plant heme peroxidases was based on function (cell wall metabolism and defence in the case of plants). The resultant phylogeny suggests that the emergence of these class III peroxidases was in response to the selective pressure placed on the ancestral plant on colonizing land (Duroux, Welinder 2003; Veitch 2004). The class III plant heme peroxidases, HRP and SBP, are the plant representatives used in this thesis.

1.1.6 *In vivo* biological function of plant peroxidases

Various physiological functions for plant peroxidases have been suggested such as the removal of H_2O_2, reduction of toxic reductants, defence against pathogens and both the synthesis (lignification) and degradation of lignin present in cells walls (Grisebach 1981; Mäder, Füssl 1982; Espelie, Franceschi, Kolattukudy 1986; Ye, Pan, Kuc 1990), (Lagrimini 1991; Dowd, Lagrimini 1997; Yoshida et al. 2003). An essential biological role for plant peroxidases is the stiffening (lignification) of cell walls resulting from the formation of bridged cell wall polymers (Fry 1986). SBP is isolated from the seed coat of the soybean plant, whereas HRP is found in the horseradish plant, with predominant isolation from the plants roots (Veitch 2004). HRP occurs as a family of isoenzymes, with up to 40 types being detected (A-E). HRP C and HRP A2 are the two most abundant isoenzymes, with the former accounting for almost half of the peroxidase activity in the root (Hiner et al. 2001; Carvalho et al. 2007). The *in vivo* physiological functions of plant peroxidases are not well known; however, their potential applications within biomedical and industrial settings are vast. Their most traditional application are in biosensor/enzyme-linked immunoassay technologies and industrial bioremediation.

1.1.7 Biomedical and industrial applications

HRP and SBP enzymes are readily used in both the biopharmaceutical/biomedical and biotechnology sectors and their applications in these sectors have been reviewed recently

(Ryan, Carolan, O'Fagain 2006). One of the most common applications of these plant peroxidases is in biosensors/enzyme-linked immunoassays. These applications have exploited the HRP function in detection of H_2O_2 and various other compounds through the use of coupled enzyme reactions (Frey et al. 2000). HRP provides a convenient and sensitive system in biosensing technologies due to its capacity to repeatedly cycle the reaction from substrate to product (Dotsikas, Loukas 2004). HRP has been the widely used as the enzyme of choice in such biosensing applications; however, SBP is fast approaching the forefront of such technologies due to its catalytic and stability properties, particularly for use in chemiluminescent enzyme-linked immunosorbent assay (CL-ELISA). HRP has been used as an enzymatic label for such techniques, however, its efficent activation (oxidation of luminol resulting in chemiluminescence) requires the addition of enhancers. These enhancers result in increased chemiluminescent intensity, which gradually degrades over time due to enzymatic inactivation by reaction by-products. SBP has recently been shown to oxidise luminol in the absence of enhancers without being inactivated by reaction products, resulting in a long-term chemiluninescent signal (Sakharov 2001; Sakharov, Alpeeva, Efremov 2006; Sakharov et al. 2010).

An important industrial application of plant heme peroxidases is in detergents and for decontamination (bioremediation) (Hiner et al. 2001; Wagner, Nicell 2002; Ryan, Carolan, O'Fagain 2006). Phenolic solutions are common in the effluents of many industries (Patterson 1985). Plant heme peroxidases, in the presence of H_2O_2, have been shown to promote the oxidation of a variety of phenolic compounds (Karam, Nicell 1997). However, such applications are limited, primarily due to the enzyme's stability in the presence of its primary substrate H_2O_2. Excess H_2O_2 is toxic to the enzyme, an action referred to as substrate suicide inactivation (Valderrama, Ayala, Vazquez-Duhalt 2002).

The major shortcoming for such applications is that they require these plant peroxidases to work efficently under harsh operational conditions. Much work has been carried out to enhance enzyme stability with the aim of preventing substrate suicide inactivation and thermal inactivation. In particular, HRP C has been studied extensively in the

improvement of its operational/catalytic performance in many sectors. Much work has been directed towards protein engineering, with particular emphasis on the site-directed mutagenesis of recombinant HRP C, which has yielded promising results in terms of enzymatic stability (Tanaka et al. 1997; Savenkova, Ortiz de Montellano 1998; Ryan, O'Connell, O'Fagain 2008; Ryan, O'Fagain 2008) Enzyme immobilisation has also been investigated in improving the enzymes operational stability. Co-immobalisation of HRP (and also SBP) with glucose oxidase enhances the plant peroxidase's oxidation stability (van de Velde, van Rantwijk, Sheldon 2001).

1.2 Protein Synthesis

Protein-coding genes/sequences in prokaryotes and eukaryotes are markedly different in several respects. Primarily, proteins in prokayotes are encoded by a continuous stretch of nucleotide triplet codons. In contrast, eukaryotic protein-coding genes consist of both coding (exons) and non-coding (introns) sections. The exonic regions consist of triplet codons. Eukaryotic proteins are encoded by genes (DNA) through a series of events, specifically transcriptions and RNA processing (nucleus) and translation (cytoplasm).

1.2.1 Transcription, translation and post-translational modifications

The inital step in the process is the generation of the short-lived intermediate transcription product, messenger RNA (mRNA). Transcription from DNA to mRNA is facilitated by RNA polymerases and transcription factors. The primary RNA transcripts consists of both intron and exon regions. RNA splicing and processing results in a mature mRNA transcript, where the intronic sections are excised from the transcript to make mRNA viable for translation.

Translation is essentially the conversion of the mature mRNA transcript into a functional polypeptide/protein. Mature mRNA is exported to the cytoplasm where it is translated by ribosomes. The ribosome first dissociates into its 40S and 60S subunits. Translation initiation factors bind to the mRNA and associate with Met transfer RNA (tRNA), which in turn bind to the 40S subunit resulting in the formation of the pre-initiation complex. This in turn binds with the 60S subunit to form the 80S ribosome complex which scans the 5' untranslated region for the start codon. Initiation factors are then released allowing translation to begin. The ribosome moves along the mRNA strand towards the 3' region, translating each codon into its corresponding amino acid and synthesising the polypeptide chain (elongation) until the final stop codon is reached (termination).

Once translation is complete, the protein undergoes a series of processing events to generate a fully folded and functional protein. These are known as posttranslational

modifications. The endoplasmic reticulum (ER) and Golgi apparatus are predominantly the intracellular locations where post-translation modifications occur. Molecular chaperone proteins assist in the folding and transport of many proteins. The addition of many chemical groups to the translated protein is essential for correct folding. The ER is the site of many modifications such as glycosylation and phospholipid biosynthesis, where sugar and lipids groups are attached during the synthesis of the protein. Disulphide bond formation is critical for stabilisation of the folded functional protein. Phosphorylation and dephosphrylation of proteins is essential regulating the activity of proteins. The Golgi is the sorting centre of the cell where proteins are targeted and packaged/stored into different vesicles and organelles.

1.2.2 Mutations

Mutations in the genomic sequence may alter the encoded protein, with three alternative outcomes; no change to the underlying protein function, loss of function or a functional change. Mutations can be induced (e.g. due to viruses, mutagenic chemical) or spontaneous (due to errors during DNA replication or meiosis). There are various types of mutations. Typically, point mutations can be classified as silent/synonymous (no change to the amino acid coded), missense (code for a different amino acid) and nonsense (can code for a premature stop codon and can result in protein truncation). Amino acid altering point muations may also be referred to collectively as non-synonymous substitutions and may have beneficial or indeed detrimental consequences for the functional protein. Insertions and deletions in the genomic sequence may also occur which disrupt the reading frame (frameshift) and mRNA splicing, altering the gene product. Chromosomal translocations and inversions have been associated with various disease states including many cancers, Down syndrome and schizophrenia (Atkins and Bartsocas 1974, Li et al 1999, Semple 2000). For further detail refer to Lewin 2003.

1.3 Molecular Evolution and Phylogenetics

The end of the 20th century heralded the rise of the field of molecular evolution. Advances in genomics and bioinformatics and the advent of whole-genome sequencing have allowed for a remarkable increase in the study of evolution from a molecular perspective. Evolutionary comparative genomics is a powerful tool for identifying the similarities (and, indeed, differences) among genomes and, in turn, provides a framework for understanding the divergence of biological function and the development of specificities.

1.3.1 Gene Duplication

Evolution is greatly influenced by gene duplication, which is believed to be the principal source of new genes (Ohno 1970). New genes may also arise from genes through recombination (Long et al. 2003). *De novo* gene synthesis from non-coding DNA is a rare source of new genes (Knowles, McLysaght 2009). Selective pressures and mutation act on existing genes over time during duplication. It has been proposed that the process of gene duplication has three possible outcomes. Gene duplicates (i) may become functionally redundant (Nowak et al. 1997), (ii) the divergence of function may occur through neofunctionalisation (Hughes 1999) where substitutions accumulate that generate a novel function or (iii) subfunctionalisation may occur, where the function of the ancestral protein is partitioned between the duplicates (Lynch, Force 2000). The events following the process of gene duplication are key to the evolution of specificity of divergent multigene families, such as the MHP.

Two of the most important mechanisms driving evolution are genetic drift and natural selection. Genetic drift is an entirely random process (random genetic drift), which generates genetic diversity whereas natural selection removes genetic diversity by selecting those mutants that have increased fitness in a given set of

conditions/environment and by removing from the population those mutants that decrease the fitness of the organism.

1.3.2 Random Genetic Drift

In the late 1960s Kimura proposed that the majority of molecuar changes in evolution are due to random genetic drift of selectively neutral mutations (Kimura 1968). Genetic drift is an important feature in evolution that leads to changes in allele frequencies from one generation to the next. This process results in the fixation or loss of alleles over time and is clearly correlated with effective population size (N_e). Population size is fundamental to the effects of genetic drift: when N_e is large, there is a small effect; however, when N_e is small, the degree of fluctuations in allele frequencies is greater. As such, the predicted time to loss or fixation is much faster in smaller populations; see Figure 1.4.

The change in allele frequencies from one generation to the next means that, for any given generation, the change in frequency is only derived from the generation immediately prior (Hartl, Clark 2007). Random genetic drift does not result in the generation of additional alleles but can result in the loss of an existing allele (allele frequency of 0); this random sampling mechanism eventually leads to loss of genetic variation. Loss of genetic variation in a continuing population is reflected by the fixation of an allele, where its allele frequency reaches 1.

The effects of drift are evident in population/genetic bottlenecks. A bottleneck occurs due to a large reduction in population size following a random external/environmental event. As a result, allelic homogenity may be increased in the population due to random genetic drift. Similar to population bottlenecks, genetic uniformity may occur when a new population arises from a small population, known as the founder effect (Mayr 1942; Provine 2004). This new population is in turn subject to the effects of random genetic drift.

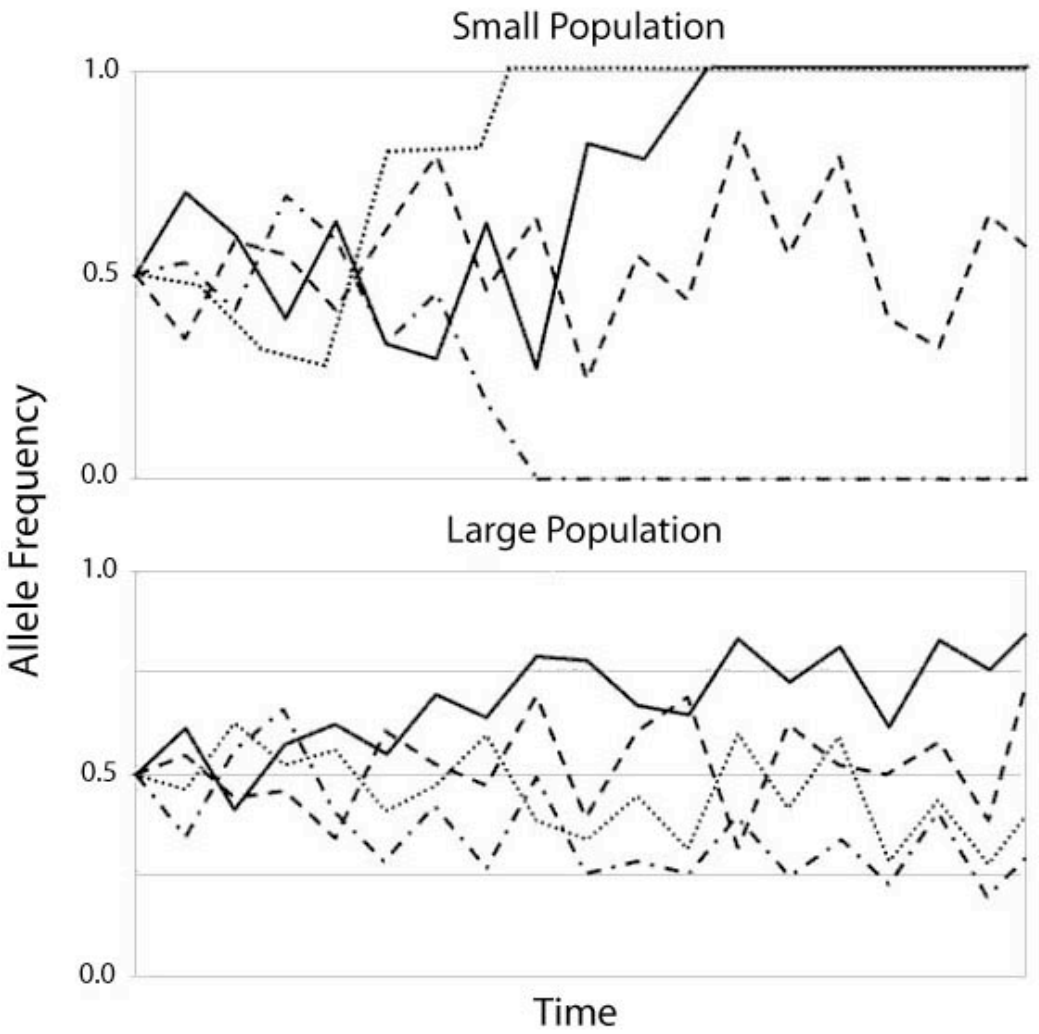


Figure 1.4. Rate of fixation of neutral alleles occuring due to random genetic drift is inversely proportional to effective population size (N_e). The effects of drift are smaller in large populatons and greater in smaller populations. Initial allele frequencies of genetic variation in both populations were equal (0.5). The smaller population reaches fixation (allele frequency of 1) or loss (allele frequency of 0) faster than the larger population. Adapted from (Johnson 2007).

As random genetic drift supplies a population with genetic diversity, natural selection acts to remove diversity by selecting for beneficial mutations and against deleterious mutations.

1.3.3 Natural Selection

The theory of natural selection is fundamental to understanding the living world. Natural selection influences the observed phenotype. The genetic variation that exists within a population due to genetic drift may not affect fitness; however, there are some mutations that affect traits/phenotype that may affect the overall chances of survival and reproduction in either a positive or negative way. Alleles favourable to the survival of an organism may be passed from one generation to the next, resulting in an increase in the frequency of the advantageous allele. Natural selection can also result in the elimination of the less fit variants from the population. Natural selection can therefore (i) remove deleterious mutations from the population (in which case it is termed "negative or purifying selection"), (ii) can fix/select advantageous mutations in the population (in which case it is termed "positive selection or adaptive evolution"), or (iii) natural selection can allow mutations to drift through the population that neither have a negative or positive effect on the organism ("neutral evolution"). Neutral evolution can be described as the accumulation of non-advantageous random neutral changes in a species; these confer no advantage or disadvantage to their host organism and are retained in a population at a rate proportional to the size of the population. Negative or purifying selection is the removal of mutations that are deleterious to the population; these are marked by a decrease in fitness such as loss of function. These mutations are weeded from the population/selected against. Positive selection/Adaptive evolution is the retention of advantageous mutations marked by an increase in fitness for the organism; these spread throughout the population. Selection of such advantageous substitutions is fundamental to the evolution of novel functions in proteins.

1.3.4 Positive Selection and Functional Divergence

Detecting positive selection is key in understanding the various mechanisms of molecular evolution, as it signifies a shift in the function of that protein (Messier, Stewart 1997; Levasseur et al. 2006b). Selective pressures acting on molecular sequence data are determined by estimating the ratio of non-synonymous mutations per non-synonymous site (D_n) to synonymous mutations per synonymous site (D_s) across the coding sequence alignment; this ratio is generally denoted as $\omega = D_n/D_s$. At the nucleotide level, non-synonymous mutations/substitutions are those that alter the amino acid coded for, whereas a synonymous substitution does not alter the encoded amino acid. These substitutions are often referred to as silent mutations, as they convey no change to the protein coded for. Non-synonymous substitutions alter the amino acid coded for and are also termed replacement substitutions; these are, therefore, visible to natural selection and are either deemed as advantageous and retained, or deleterious and removed from the population.

The ratio of D_n/D_s fall into three categories: A $D_n/D_s > 1$, indicates that adaptive evolution is frequent, reflecting a high rate of functional protein divergence through directional selection. The opposite is true for purifying selection where $Dn/Ds < 1$. When there is no difference between the rate of silent and replacement substitution D_s and D_n are equal and the ratio $D_n/D_s = 1$, i.e. neutral evolution. In this thesis Dn/Ds is referred to as ω as it is derived in a maximum likelihood framework.

The calculation of D_n/D_s is explained in Figure 1.5. From (i) it is clear that there are 2 non-synonymous substitutions and 4 nucleotide substitutions in total, meaning that 2 of which were synonymous substitutions. In (ii), the total numbers of non-synonymous and synonymous nucleotide sites are calculated for the same pairwise alignment. At site 1, A is present in both sequences, in the codons ACT and ACG (coding for Thr). If A at the first position in this codon is changed to C, T or G, the encoded amino acid is altered (Pro, Ser, Ala respectively). As such, site 1 is a non-synonymous site (a score of 1 is given to non-synonymous sites and 0 to synonymous sites). At site 2, in the genetic

```
(i) Sites:        123 456 789 111 111 111
                              012 345 678

Pairwise          ACT CCG ACT AAA GCG CCC
alignment          *       *       *       *
                  ACG CCG CCT AAG GCG CTC

Amino             T   P   T   K   A   P
acids                     *           *
                  T   P   P   K   A   L

Non-syn substitutions: 2
Syn substitutions: 2
```

```
(ii) Sites:       123 456 789 111 111 111
                              012 345 678

Non-syn           110 110 110 11⅔ 110 110

Syn               001 001 001 00⅓ 001 001

Non-syn sites: 12.67
Syn sites: 5.33
```

(iii) D_n = No. non-syn substitutions / No. non-syn sites

$D_n = 2 / 12.67 = 0.158$

D_s = No. syn substitutions / No. syn sites

$D_s = 2 / 5.33 = 0.375$

$D_n / D_s = 0.421$

Figure 1.5. Calculating D_n/D_s. A Pairwise alignment of homologous coding DNA sequences can be seen in (i) with each site in the alignment allocated a number (Sites). The asterisks (*) indicate the nucleotide sites at when the two sequences differ (i.e. where substitutions have occurred). Their encoded amino acids (AA) can be seen below the pairwise alignment. Here the asterisks indicate the number of nucleotide substitutions that have altered the encoded protein (non-synonymous substitutions). In (ii), the total numbers of non-synonymous and synonymous nucleotide sites are calculated for the same pairwise alignment. D_n and D_s and their ratio are calculated in (iii). Non-syn: non-synonymous, Syn: synonymous.

code, no matter what codon is present, altering the second position in that codon always alters the encoded amino acid, thus position 2 is always a non-synonymous site. Moving to site 3, the third position in the codon, the codons present is ACT and ACG, again both coding the same protein, Thr. If site 3 was altered to A or C, the encoded amino acid is still Thr, therefore this is a synonymous site (a score of 0 is given to non-synonymous sites and 1 to synonymous sites). Sites are designated as non-synonymous or synonymous in this way moving across each site in the alignment. Moving to site 12, position 3 in the codons AAA and AAG, both positions differ but the encoded amino acid (Lys) does not alter. If position 3 was changed to C or T, the amino acid coded for changes (both Asp). This means that position 3 in this codon represents a two-fold degenerate site and as such is considered 2/3 non-synonymous and 1/3 synonymous. Calculating the total number of non-synonymous and synonymous sites gives 12.67 and 5.33 respectively. D_n and D_s and their ratio are calculated in (iii). D_n is the number of non-synonymous substitutions per non-synonymous site (0.158) and D_s is the number of synonymous substitutions per synonymous site (0.375). From this the D_n/D_s ratio can be calculated (0.421). In this example, $D_n/D_s < 1$, indicating negative/purifying selection - that is some non-synonymous mutations are deleterious and the rest are neutral.

This standard method for detecting what selective pressures are at play can be performed irrespective of the evolutionary relationship of the observed data. Pairwise sequence comparisons of these rates of substitution can be calculated. Signatures of positive selection can be identified in specific clades/branches on a phylogeny, making it possible to pinpoint the specific sites that may be responsible for the observed functional shift in that lineage (Messier, Stewart 1997; Levasseur et al. 2006b). The first analysis of this type was carried out on primate lysozymes (Stewart, Schilling, Wilson 1987; Messier, Stewart 1997). Lysozymes are important in bacterial defence. Their analysis showed episodic positive selection in the lysozyme protein in ancestral colobine lineage (unique foregut-fermenting old world primates) followed by a short period of negative selection, reflected by greater levels of nonsynonymous site differences than synonymous site differences for the colobines. The positively selected residues identified in the foregut-fermenting primates are believed to be associated with the

enzymes ability to function in the hostile conditions of the stomach, with these enzymes operating at a lower pH than other lysozymes.

Use of the more sophisticated maximum likelihood (ML) and Bayesian based methods enables the incorporation of the evolutionary relationship of the observed data. These methods are more robust to detecting what selective pressures are acting on the data, and also to what degree adaptive evolution has influenced particular lineages (Yang 1997). Such approaches consider the complexity of evolutionary models, the phylogenetic relationship and the data. These methods allow for the use of improved evolutionary (codon) models. The advantage of this over previous standard approaches are that important substitution properties of protein coding sequences other than the D_n/D_s ratio are considered such as transition to transversion rate ratios and codon usage frequency and that substitutions are considered one codon site at a time. Ignoring such properties by applying an *ad hoc* treatment to data and making simple assumptions about the evolutionary process can lead to estimation errors. Over estimations of synonymous site and in turn D_s may be due to the increased likelihood of synonymous transitions at the third codon position rather than transversions (Li, Wu, Luo 1985), hence an underestimation of D_n/D_s. Codon usage biases result in varied or unequal substitution rates among the codons and can also lead towards biased estimates if overlooked. The most extensively used software for phylogeny-based detection of positive selection is Yang's phylogenetic analysis by maximum likelihood (PAML) package, which incorporates various models of codon evolution (see Chapter 2) (Yang 1997; Yang et al. 2000). This method allows for the estimation of various selective pressures among sites and across lineages (clades). As such, selective pressures acting on gene duplicates may be detected. Under ML, the proportion of sites falling into each ω-value category is estimated. A "measure of fit" of the tested models to the data is determined by performing likelihood ratio tests (LRTs) by comparing the likelihood score of a simple model with the likelihood score of its more parameter-rich extension. Empirical Bayesian methods are implemented to determine the posterior probability of the identified sites being in their respective category. Sites with a high posterior probability coming from the class where $\omega > 1$ are inferred as positively selected sites. The

Bayesian estimations use the ML parameters of proportion and corresponding D_n/D_s ratio for a particular site as priors for calculating the posterior probability at that given site. Naïve empirical Bayes (NEB) estimations use these ML parameter estimates with the assumption that they are correct without accounting for error, which is not always the case, and can potentially result in false positive inferences of sites under positive selection. NEB is particularly sensitive to error in small datasets where ML estimates may have large sampling errors (Anisimova, Bielawski, Yang 2002). Bayes empirical Bayes (BEB) approaches have been developed to account for these uncertainties, however, it has not been implemented for all models of codon evolution (Yang, Wong, Nielsen 2005).

1.3.5 Phylogenetics

Identifying and understanding genome similarities and differences by means of evolutionary analysis underpins the study of the evolutionary relatedness amongst genomes, commonly referred to as molecular phylogenetics.

The evolutionary history of genetic information (or characters) from species is reconstructed in the form of a gene tree whereas a species phylogeny depicts the evolutionary relationships amongst species. A gene tree is reconstructed from a single homologous gene across a variety of taxa, the relationships depict the history of the gene over the evolutionary timescale. On the other hand, a species phylogeny is reconstructed using large amounts of homologous sequence data from the species of interest and identifying the most highly supported consensus tree (using, for example, concatenated alignment or supertree methods). In the case of a species phylogeny, the relationships depict the divergence of species over the evolutionary timescale. Gene trees do not always recapitulate species trees; this is because every gene has a unique evolutionary history due to (for example) mutational rate variation (rate heterogeneity) along genes and in particular species, and/or gene gain and loss. Therefore, the branching pattern of a gene tree is expected to converge with the species tree from which the genes have been

sampled *provided* the data is orthologous and reasonably homogeneous in terms of rate variation.

Using large amounts of data and ML and Bayesian methods of phylogeny reconstruction, the evolutionary relationship of mammals has been fully resolved (Murphy et al. 2001) (see Figure 1.6 for summary). Since its publication there have been a small number of challenges to this topology. To date, however, all have been found to be due to artefacts in data which have been highlighted by careful examination (Cannarozzi, Schneider, Gonnet 2007; Lunter 2007). For any given gene/protein family, cases of rapidly evolving lineages, or of gene gain and loss, can, therefore, be identified by comparing the mammalian gene tree to the fully resolved true branching pattern of the represented species on the species phylogeny (Page, Holmes 1998). By using only completed genomes, we can be confident that, on reconciling the relationships between gene trees and their corresponding species trees, we can identify the pattern of gene loss and duplication and understand any incongruence existing in the data (Page 1998; Page, Holmes 1998; Page, Cotton 2002).

Several approaches may be employed to reconstruct the phylogenetic relationships of molecular sequences. These methods fall into two general categories: distance-based methods and character-based methods.

The former is the more simplistic method, where a distance-matrix is calculated for all pairwise combinations of all sequences/taxa present and, based on this matrix a phylogenetic tree is computed with branch lengths closely resembling the distance between sequences/taxa. Minimum evolution (ME) inference transforms the data into pairwise distances. This method computes the sum branch lengths from the pairwise distances, where the estimated ME phylogeny is the one that requires the minimum length. These branch lengths are often estimated using least-square methods. However, ME inferences using least-squares has been found to be less robust than other distance-based methods, such as neighbour joining (NJ) (Willson 2005). The most popular distance-based algorithmic methods are unweighted pair-group method with arithmetic

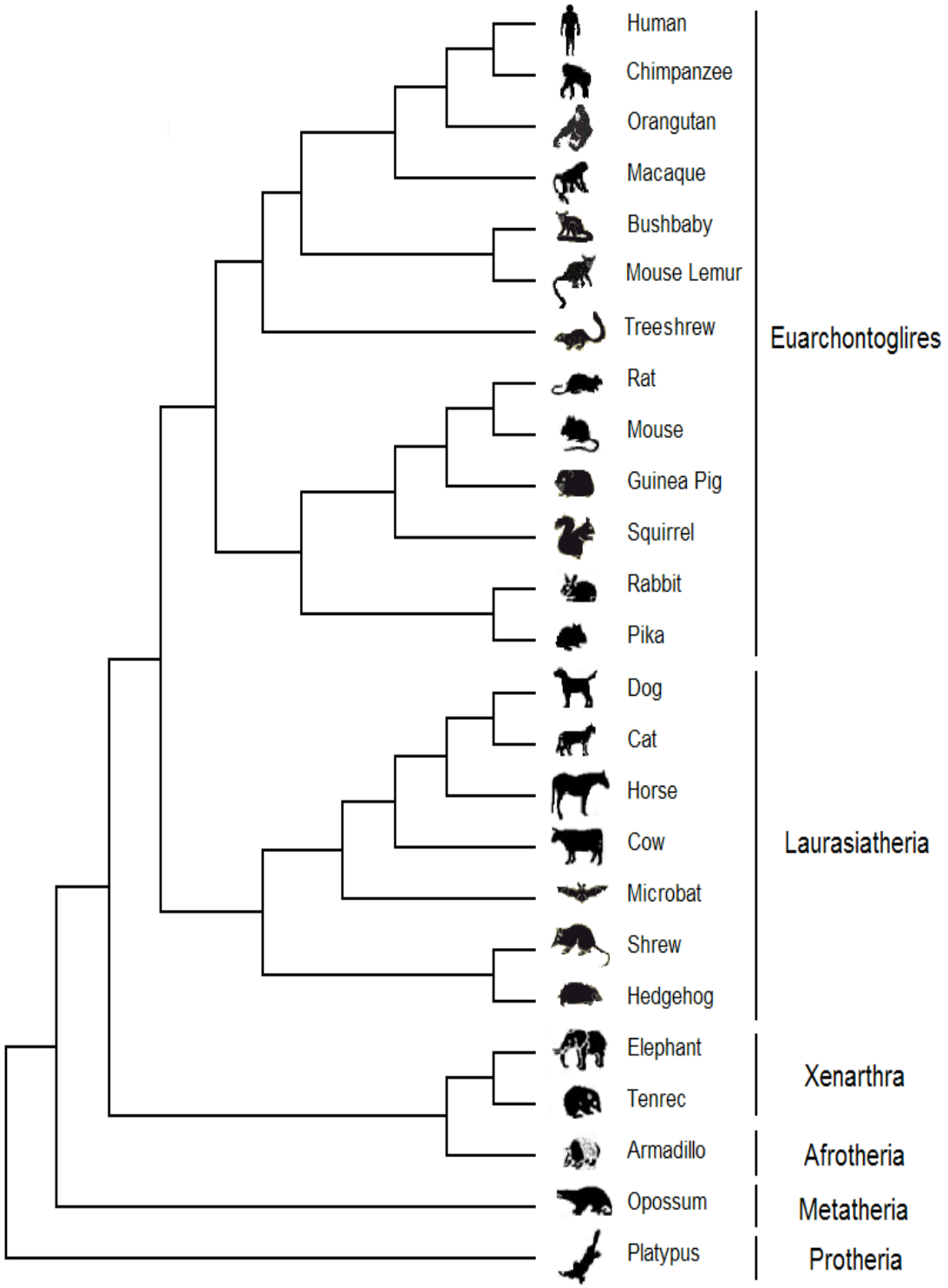


Figure 1.6. Resolved phylogeny of a selection of fully sequenced mammalian genomes. Adapted from Murphy *et al.* 2001 and (Benton, Donoghue 2007).

mean (UPGMA) and NJ (Sokal, Michener 1958; Saitou, Nei 1986; Saitou, Nei 1987). Both methods involve a step-wise manipulation of the distance matrices calculated from the dataset. Although these methods of reconstruction are fast, compressing sequence data into distances can result in the loss of information and the estimated evloutionary distance/change may be inaccurate. Once the information pertaining to specific characters are compressed into distance, an overall estimate of the resultant tree and data is determined and evolutionary changes/information about specific characters is lost. Distances can be estimated from characters but not *vice versa*.

Character-based methods are more computationally intensive, as they consider each individual nucleotide or amino acid position in the multiple sequence alignment (MSA) directly and reconstruct the evolutionary tree best suited to the information collected. The most common tree searching character-based methods are parsimony, ML and Bayesian methods (Fisher 1912; Edwards, Cavalli-Sforza 1963; Edwards, Cavvalli-Sforza 1964; Neyman 1971; Felsenstein 1978; Swofford 1993; Ronquist, Huelsenbeck 2003; Keane, Naughton, McInerney 2007). Under maximum parsimony, the estimated phylogeny is the one that required the minimum amount of evolutionary change to explain the observed data. The most robust of the discrete methods are ML and Bayesian inference; however, the major shortcoming of these favoured methods is the computational power required. ML phylogeny estimations are carried out on the data given a model (implemented in programmes such as MulitPhyl (Keane, Naughton, McInerney 2007)). The likelihood measurement (L) used to select between various alternative trees is the probability (P) of observing the data (D) given the tree (τ), branch lengths (ν) and the substitution model (evolutionary process) (θ) expressed as

$$L = P(D \mid \tau, \nu, \theta) \quad Eq. \quad 2$$

Bayesian phylogeny inference is a variation of ML, in which Bayes' theorem is applied, where the resultant tree is one which has the greatest probability given a set of priors (priors are some idea about how the data has evolved) and the data (D) (implemented in programmes such as MrBayes (Ronquist, Huelsenbeck 2003)). Such information is

expected to enhance/strengthen inference (Huelsenbeck et al. 2002). The priors in Bayesian inference are the tree (τ), branch lengths (ν) and the substitution model (evolutionary process) (θ), these priors all constitute the model (M). The posterior probability of the model given the data ($P(M|D)$) is the probability of the data given the model ($P(D|M)$) (*likelihood*) times the prior probability of the model ($P(M)$) (*prior*) divided by the probability of the data ($P(D)$) (*normalising constant*), expressed as

$$P(M \mid D) = \frac{P(D \mid M) \times P(M)}{P(D)} \qquad Eq. \quad 3$$

The use of amino acids over nucleotides is favoured in phylogeny reconstruction. Nucleotide datasets are more prone to compositional bias due to differences in codon usage. The reason we use amino acids is that there are more possible states (20) (rather than 4 with nucleotides); this increases the power to infer ancestral conditions and, therefore, phylogenetic reconstruction. Also the use of amino acids may help circumvent the problem of nucleotide compositional bias. Several substitution models (20x20 matrices) have been developed to describe the rate of change between amino acids; these models have been based on amino acid frequency in the environment, physicochemical properties and genetic distances etc (e.g. Dayhoff and Jones, Taylor Thornton (JTT)) (Dayhoff, Schwartz, Orcutt 1978; Jones, Taylor, Thornton 1992). Selection of an appropriate model is fundamental to the estimation of accurate evolutionary relationships, as can be seen from the likelihood and posterior probability expressions above. A number of tests can be performed to determine the optimal model of evolution, all involve the direct comparison of the scores generated by fitting alternative models to the data (Posada, Crandall 1998; Posada 2003; Keane et al. 2006; Posada 2006). These include the likelihood ratio test (LRT), the Akaike information criterion (AIC) and Bayesian information criterion (BIC) methods amongst others (Akaike 1974; Schwarz 1978; Cao et al. 1997). The LRT estimates a likelihood score that reflects the "measure of fit" between the model and data through a series of pairwise comparisons between different models. The more complex methods, AIC and BIC, focus on the different models independently and not in a pairwise manner, considering both the goodness of fit

and complexity of the model. Programmes such as ProtTest and MODELGENERATOR have been developed that implement such tests in determining the substitution model that best fits the data (Abascal, Zardoya, Posada 2005; Keane et al. 2006).

Phylogeny reconstruction is difficult for a number of reasons outlined below and requires careful analysis and knowledge of the data. Even though amino acids are favoured over nucleotides in phylogeny reconstruction, to circumvent the base composition bias prevalent with nucleotide data, amino acid composition bias can also be present in the data. The amino acid content of proteins is restricted by the base composition of their cognate genes. Like base compositon bias, phylogenetic analysis may be incorrect due to amino acid content and may indicate false/lack of homology. Inadequate signal is another difficulty with phylogeny reconstruction, particularly if one is dealing with closely related species or recent speciation events. A thorough analysis of the data present in the dataset and in the associated genomes is necessary to ensure that paralogs are not mistaken for orthologs. These are two different types of homologous genes. Orthologous are genes that are separated by a speciation event whereas paralogs result from a gene duplication event. Distinguishing between the two is critical for the reconstruction of species trees and indeed gene phylogenies. The final difficulty with phylogeny reconstruction is by far the most challenging and perhaps the most prevalent, namely the effect of long branch attraction (LBA). LBA effects are seen in phylogenies as a direct consequence of variable evolutionary rates (heterogeneity in the data). This LBA artifact is responsible for the inference of misleading phylogenies, as it leads to the clustering together of fast evolving sequences towards the base of the phylogeny, regardless of their true evolutionary relationship (Felsenstein 1978). The basal positioning of species on a tree may be influenced by their germ line generation times. Under the so called 'Generation Time Hypothesis', the rate of mutational divergence between species can be attributed to the germ line generation time of the species. The germ line generation time is the number of germ line cell divisions that occur per unit time and under this hypothesis it is negatively correlated with the mutational/substitution rate. We would expect that those lineages that have a larger generation time have a lower substitution rate. Conversely, more rapidly evolving lineages accumulate a greater

number of mutations than species with shorter generation times due to greater number of germ line replications per year (Laird, Mcconaug.Bl, Mccarthy 1969; Kohne 1970). Species with shorter generation times have a greater genetic diversity than those species with longer generation times; see Figure 1.7 for the germ line generation times of a selection of completed mammalian genomes.

In the MHP dataset studied in this thesis (Chapter 2), the following species have been included: human, chimp, macaque, rat, mouse, dog, cow and opossum. Therefore, there are a mixture of germ-line generation times (varying ages of reproductive maturity from approximately 4 months to 200 months, see Figure 1.7) and, hence, a variation in substitution rates amongst lineages. This phenomenon can be seen by a mixture of branch lengths among evolutionary lineages which is indicative of the LBA effect. The results of the analysis follow (Chapter 2). LBA can be minimised or perhaps even overcome entirely by the addition of more taxa to the dataset, spanning the interval between fast and slow evolving species, with the aim of breaking up long branches (e.g. Treeshrew on Figure 1.6 breaks the branch between the primates and the rodentia). This has been seen in the analysis of early animal evolution, where the debate lingered for some time as to whether the correct grouping of animals was Coelomata (the presence or absence of a coelem) or Ecdysozoa (the presence or absence of skin shedding) (Mader 1993; Aguinaldo et al. 1997). The most recent datasets used to address this question, incorporate multiple species of intermediate evolutionary rate and more data than has been previously applied to this problem. Combined, the sampling and data size have converged on support for the Ecdysozoa and rejection of the Coleomata as a hypothesis for early animal evolution (Mushegian et al. 1998; Philippe, Lartillot, Brinkmann 2005; Holton, Pisani 2010)

Alternatively the effects of LBA can be minimised by simply using improved models of substitution that account for rate heterogenity and amino acid composition bias in molecular evolution (Philippe et al. 2005a; Lartillot, Brinkmann, Philippe 2007). Finally, it has been posed that the phylogeny can be resolved by concentrating on the slower

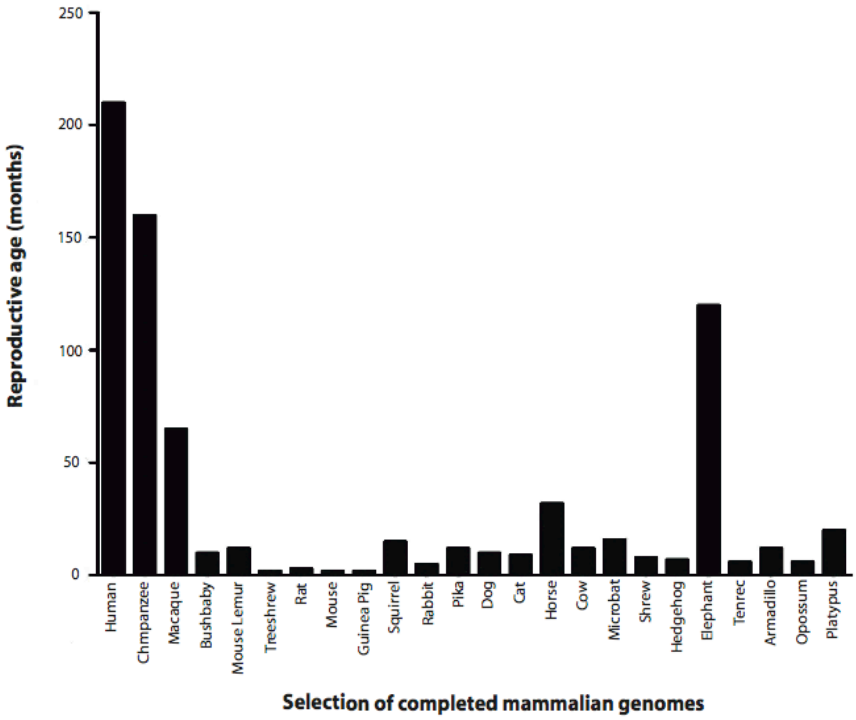


Figure 1.7. Germ-line generation times of a selection of completed mammalian genomes. Approximate age of reproductive maturity in months. Adapted from (Morgan et al. 2010).

evolving sites (*Slow Fast method*) of the homologous alignment, see Figure 1.8 (Gribaldo, Philippe 2002).

Reconstructing the evolutionary histories of genes and protein families has many potential applications. An estimated phylogeny can help in determining what selective pressures are influencing the observed relationships and, thus, pinpoint specific sites associated with positive selection/functional divergence. A resolved phylogeny can also allow for the reconstruction of ancient protein sequences, making possible the resurrection of such ancestral states which can provide an insight into the evolution of protein function.

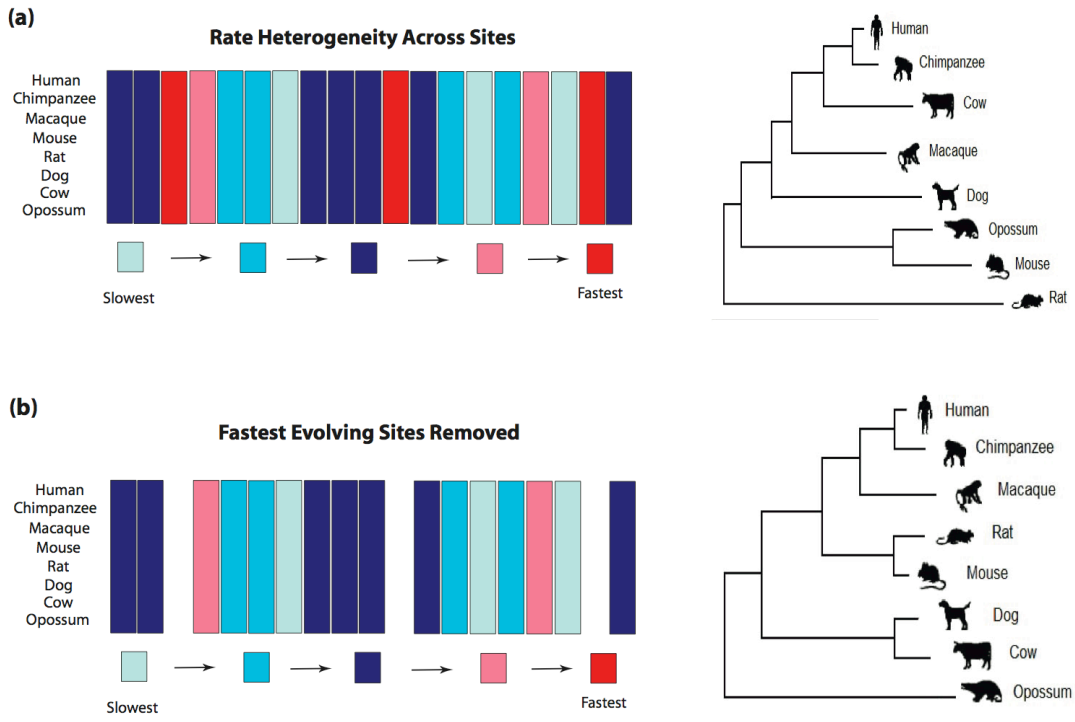


Figure 1.8. Removal of fast evolving sites to minimise long branch attraction. The rate heterogeneity varies across the sites of a multiple sequence alignment and can result in the inference of the incorrect phylogenetic relationship due to long branch attraction (LBA) (a). By removing the most rapidly evolving sites and concentrating on the slow evolving sites (*Slow Fast method)* the effect of LBA and result in the reconstruction of the true evolutionary relationship (b).

1.4 Ancestral Protein Resurrection (Paleomolecular Biology/Biochemistry)

1.4.1 The rise of paleomolecular biology/biochemistry

In 1963 two scientists, Linus Pauling and Emile Zuckerkandl, proposed that the evolutionary history of proteins (and, therefore, of organisms and species) was written in the molecules (we now know that the "molecules" referred to are the genetic sequences of proteins and organisms) (Pauling, Zuckerkandl 1963). This marked the beginning of the field of molecular evolution. The advent of molecular evolution and phylogeny reconstruction has provided powerful tools for understanding biology: with the ability to map molecular changes over time (in some cases millions of years) we can glean a greater understanding of living systems. The concept of studying the biochemical properties of extinct proteins was initiated in the 1980's in work by Benner (Benner 1988). With a fully resolved evolutionary relationship of a family of proteins, the resurrection of ancestral sequences is possible. This type of analysis is known as paleobiochemistry or paleomolecular biochemistry.

In the previous section (Section 1.2) the concept of reconstructing the phylogenetic history of evolutionarily related proteins (homologs) was described. To reconstruct the phylogeny of a set of proteins the algorithm (any of those in existence) must first use the extant samples to reconstruct the ancestor at every node in the phylogeny. As one moves further back in time, the reconstructed nodes are used to reconstruct more further nodes etc. As one reconstructs the phylogeny, one therefore moves from extant to extinct nodes, and the sequence for every extinct node is reconstructed. The process is outlined in detail in Section 1.3.2 below. By recreating ancestral states, it is possible to determine from what state the extant proteins have evolved, allowing for direct study of the emergence of functions of modern-day proteins and a better understanding of evolutionary processes (Chang, Donoghue 2000; Cai, Pei, Grishin 2004; Thornton 2004; Benner, Sassi, Gaucher 2007). After approximately two decades, Pauling and Zuckerkandl's vision of studying evolution using sequences was finally realised due to

advances in technology. It is only the last quarter of a century that has witnessed the rise in paleomolecular biology. Since the late 1980's, only approximately two-dozen reports on molecular resurrections have been reported: see Table 1.2 below (Benner, Sassi, Gaucher 2007).

Early paleomolecular biochemistry analyses focused on the evolution of digestive enzymes (Benner 1988; Stackhouse et al. 1990; Jermann et al. 1995). Following these initial ancestral resurrections, much controversy over the concept of paleomolecular biology was evident, as expressed by Nicholas Wade's article, "Method & Madness; Dead Sure", in the *New York Times Magazine*, where he concluded that "the stirring of the ancient artiodactyls ribonuclease is a foretaste of biology's demiurgic powers. It may well prove best to keep resurrection an unroutine event" (Wade 1995). With the turn of the 21st century, ancestral resurrections have been used in the study of the evolution of several biomolecular systems, including the adaptive evolution of visual pigments and steroid hormone receptors (Chang 2003; Thornton, Need, Crews 2003; Chinen, Matsumoto, Kawamura 2005a; Chinen, Matsumoto, Kawamura 2005b; Benner, Sassi, Gaucher 2007), with findings supportive of Benner's idea that "the past is the key to the present" (Benner 2002).

Not only can paleomolecular biochemistry provide insights into mechanisms of evolution, but it can also potentially harness biological functions/processes that may have been lost over time. Rediscovery of such "lost" functions may be beneficial in biomedical or industrial applications today.

1.4.2 Resurrecting ancestral proteins

The strategy for resurrecting ancient genes/proteins involves six main steps (Thornton 2004). In brief: (i) a highly supported evolutionary relationship of a dataset of interest is inferred using methods such as maximum parsimony or ML, (ii) using this phylogeny and corresponding dataset, the protein sequence at the ancestral node(s) of interest are reconstructed (by ML for example: see Section 1.3.3), (iii) using this hypothetical

Table 1.2: Examples of ancestral resurrections from literature to date.

Extant Genes	Ancestral Gene Resurrected	Approximate Age (million years)	Reference
Digestive ribonuclease	Ancestor of buffalo and ox	5	(Benner 1988; Stackhouse et al. 1990)
Digestive ribonuclease	Digestive RNases in the first ruminants	40	(Jermann et al. 1995)
Lysozyme	Ancestral bird lysozyme	10	(Malcolm et al. 1990)
L1 retrotransposons in mouse	Ancestral rodent retrotransposon	6	(Adey et al. 1994)
Chymase proteases	Ancestral ortholog in LCA of mammals	80	(Chandrasekharan et al. 1996)
Sleeping Beauty transposon	Active ancestral transposon from fish	10	(Ivics et al. 1997)
Tc1/mariner transposons	Ancestral paralog genomes of eight salmonoids	10	(Ivics et al. 1997)
Immune RNases	Ancestral otholog LCA of higher primates	31	(Zhang, Rosenberg 2002)
Antiretroviral proteins	Ancient primate antiretroviral proteins	< 25	(Goldschmidt et al. 2008)
Pax transcription factors	Ancestral paralog	600	(Sun et al. 2002)
SWS1 visual pigment	Ortholog in LCA of bony vertebrates	400	(Shi, Yokoyama 2003)

Extant Genes	Ancestral Gene Resurrected	Approximate Age (million years)	Reference
Vertebrate rhodopsins	Archosaur opsins	240	(Chang et al. 2002)
Fish opsins (blue, green)	Fish opsins	30-50	(Chinen, Matsumoto, Kawamura 2005b)
Steroid hormone receptors	Ancestral paralog	600	(Thornton, Need, Crews 2003)
Yeast alcohol dehydrogenase	Enzyme at origin of fermentation	80	(Thomson et al. 2005)
Green fluorescent proteins	Ancient fluorescent proteins	ca. 20?	(Ugalde, Chang, Matz 2004)
Isocitrate dehydrogenase	Ancestral eubacteria	2500	(Zhu, Golding, Dean 2005)
Isopropylmalate dehydrogenase	Ancestral archaebacteria	2500	(Miyazaki et al. 2001)
Isocitrate dehydorgenase	Ancestral archaebacteria	2500	(Iwabata et al. 2005)
Elongation factors	LCA of eubacteria	3500	(Gaucher et al. 2003)
Elongation factors	Ancestral eubacteria	500 - 3500	(Gaucher, Govindarajan, Ganesh 2008)

Note: Ages are approximate and in some cases conjectural. Shaded in grey are those examples that have inferred and synthesised complete ancestral states and not isolated point mutations. LCA: last common ancestor. Adapted from Benner, Sassi and Gaucher 2007.

sequence of amino acids, the corresponding nucleotides are inferred based on the optimum codon usage of the host system, (iv) this sequence is then synthesized by either stepwise PCR (or, site-directed mutagenesis may suffice if only a few changes exist between the ancestral and extant sequences), (v) the assembled ancestral gene is then cloned into an expression vector that will allow high-level expression in the host system of interest. The resultant plasmid is then transformed or transfected into a suitable host system, with subsequent expression of the ancestral protein, (vi) the expressed ancestral protein is purified (if necessary) and its biochemical functions are characterised using appropriate experimental techniques. See Figure 1.9 for description of the process of ancestral sequence reconstruction.

1.4.3 Reconstructing ancestral sequences

There are four main approaches to generating an ancestral sequence: (1) consensus approach, (2) maximum parsimony, (3) ML, and (4), Bayesian inferences. A small number of studies have inferred ancestral states using the consensus approach. This method is performed irrespective of the phylogenetic relationship of the data and assumes that the most frequent extant state in a dataset is that of the ancestral state. This approach is sensitive to error when using very divergent datasets but since the realisation of the limitations of this approach it has not been popular (Adey et al. 1994; Ivics et al. 1997).

The phylogeny-based methods, maximum parsimony, ML and Bayesian, all consider the evolutionary relationships in the dataset when estimating ancestral sequences. The leaves/terminals on a tree represent the extant sequences, whereas the internal nodes represent an ancestral state. These internal nodes can represent one of two events in the history of the protein, (i) a gene duplication event, or (ii) a speciation event. A highly supported phylogeny is fundamental to accurate estimation of ancestral states.

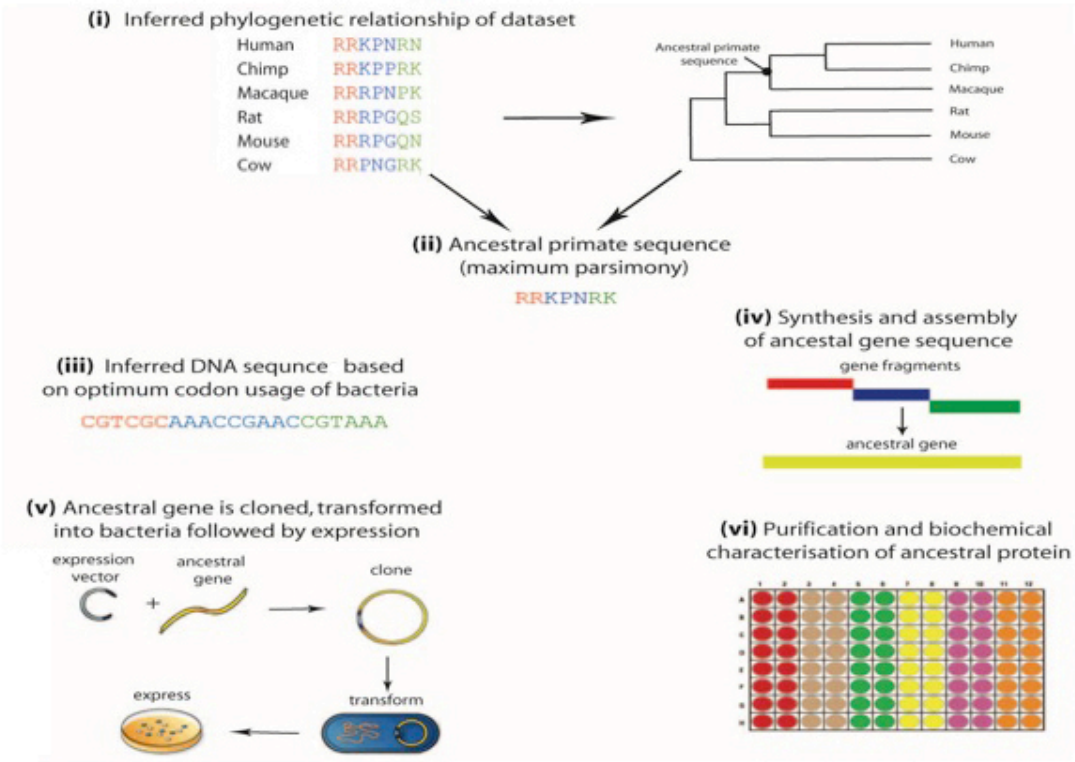


Figure 1.9. Ancestral protein resurrection. Successful resurrection of ancient proteins involves six main steps: (i) inferring the evolutionary relationship of the given dataset, (ii) using this phylogeny and corresponding dataset, reconstruct the protein sequences at ancestral node of interest (by maximum parsimony for example), (iii) using this hypothetical sequence, infer the corresponding DNA sequence based on optimum codon usage of host system, (iv) this ancestral nucleotide sequence is synthesised followed by (v) cloning and expression of ancient protein and (vi) purification (if necessary) and biochemical characterisation.

Maximum parsimony methods infer ancestral sequences assuming the minimal amount of evolutionary change and equal mutational rates across lineages (implemented in programmes such as PAUP (Swofford 1993)). Although maximum parsimony was a welcomed advance to the consensus approach, it too has limitations. It is a reasonably accurate method for reconstructing ancestral states of extant sequences that are closely related to each other (Zhang, Nei 1997). However, with more divergent sequences that have varying rates of change and, hence, varying branch lengths across the phylogeny, parsimony does not function well (Benner, Sassi, Gaucher 2007).

ML and Bayesian methods are more robust for inferring ancestral states over maximum parsimony. Likelihood/Bayesian reconstruction methods consider varying rates of sequence evolution across lineages (implemented in programmes such as PAML and MrBayes (Yang, Kumar, Nei 1995; Yang 1997; Ronquist, Huelsenbeck 2003; Yang 2007) This allows for more accurate estimations of ancestral states for more divergent extant sequences because the evolutionary process is considered. For each character in the sequence, the most probable ancestral state inferred by ML is determined statistically (Bayesian posterior probabilities). Using ML approaches, one uses the phylogeny which is generated using any of the methods deemed suitable as outlined in section 1.2.4. This phylogeny is known as *a priori* and is assumed to be correct. As such, any uncertainty over the tree is overlooked. Bayesian inferences of ancestral states consider this uncertainty by estimating the likelihood of each ancestral state for an array of possible phylogenies. This approach is extremely computationally intensive. It has recently been shown that ML inferences of ancestral states are accurate and that Bayesian approaches to incorporate phylogenetic uncertainty are not necessary when reconstructing ancestral sequences (Hanson-Smith, Kolaczkowski, Thornton 2010).

1.4.4 Ancestral resurrections in practice

There has been on average almost one report of ancestral resurrections each year since Benner and Stackhouse and colleagues resurrected the first two ancient proteins in 1988

and 1990 respectively (Benner 1988; Stackhouse et al. 1990). Improvements in computational techniques have allowed for relatively accurate estimations of ancestral states. The result has been an increase in the number of studies in the field of paleomolecular biochemistry. The results of these investigations (Table 1.2) provide invaluable insights into the dynamics of evolution. Their findings have been extensively reviewed by Thornton and Benner and cover both plants and animals (Thornton 2004; Benner, Sassi, Gaucher 2007). Below, two of the most recent studies of ancestral proteins are highlighted, involving several ancestral sequences across different lineages that have been resurrected and characterised. The estimated age of the ancestral proteins resurrected spans from under 14 million years (MY) to over one billion years.

1.4.4.1 Paleotemperatures

In an extension to a previous study (Gaucher et al. 2003), Gaucher and co-workers resurrected over a dozen ancestral proteins related to the extant thermally diverse elongation factor (EF-Tu) proteins that are crucial in the regulation of protein synthesis (Gaucher, Govindarajan, Ganesh 2008). These studies have resurrected the most ancient proteins to date, with ancestral proteins that are estimated to have existed over one billion years ago. The ancestral protein states were reconstructed by ML and Bayesian estimations implemented in the PAML package (Yang 1997; Yang 2007). Gene synthesis was achieved by step-wise overlap extension PCR. The resultant ancient genes were cloned, expressed and purified to assess their thermal capacities. From their findings it is evident that ancient bacteria were more thermophilic with respect to their extant counterparts. The thermal stabilities of these ancient proteins correlate to the temperate climate of the Precambrian eon (~ 0.5 – 3.5 billion years ago) (Gaucher et al. 2003; Thornton 2004; Benner, Sassi, Gaucher 2007; Gaucher, Govindarajan, Ganesh 2008; Gouy, Chaussidon 2008).

1.4.4.2 Ancestral immunity

In 2008, Goldschmidt *et al.* reported the antiretroviral activity of several ancient primate TRIM5α proteins (Goldschmidt et al. 2008). TRIM5α is a member of the tripartite motif (TRIM) family (Reymond et al. 2001). It has the ability to restrict various lentiviral infections such as human immunodeficiency virus (HIV) and simian immunodeficiency virus (SIV). The degree of its antiretroviral activity towards these viruses varies across the primate lineage. Goldschmidt *et al.* resurrected five ancestral TRIM5α proteins across the primate phylogeny. The most ancient resurrection was the old world primate ancestor (~ 25 MY ago). ML and Bayesian methods inferred the ancestral state at each of the five nodes. The ancestral sequence was resurrected by sequential site-directed mutagenesis of an extant human TRIM5α clone with subsequent generation of cells stably expressing the TRIM5α variants. They found that the restriction efficiency against HIV-2 and SIV variants was greater for extant TRIM5α than for the resurrected antiretroviral proteins. However, the most ancient TRIM5α protein had a high antiretroviral restriction capacity to HIV-1 relative to that of modern TRIM5α, with HIV-1 restriction diminishing across the phylogeny towards tips/leaves. Positively selected sites (as described in section 1.2.3) were also identified in human TRIM5α. Their positions in the old world primate ancestral sequence were mutated to one of two possibilities, either the ancestral state of (i) the last common ancestor of the African apes or (ii) the human/chimpanzee ancestor. The result was a reduction in the restriction capacity of the old world primate TRIM5α to HIV-1. Goldschmidt *et al.* (2008) suggested that these varying restriction capacities to lentiviral infections might be attributed to lineage-specific pandemics ultimately altering intrinsic defence mechanisms.

These findings, together with those reviewed previously (Thornton 2004; Benner, Sassi, Gaucher 2007), highlight the increasing importance of paleomolecular biochemistry in increasing our understanding of the evolutionary processes of biological function.

1.5 Aim of thesis

Identifying and understanding the similarities, and indeed differences, between sequences underpins the study of the molecular evolution of genes and genomes, known as molecular phylogenetics. The mutational events following gene duplication events are fundamental to the evolution of functional specificity in divergent multigene families. Positive selection/Adaptive evolution is the retention of beneficial mutations, i.e. those marked by an increase in fitness, in the population. Detecting positive selection is fundamental to understanding the evolution of protein functional shift. A combination of phylogeny and molecular evolutionary theory can pinpoint specific residues associated with positive selection (functional divergence) and can also enable us to reconstruct ancient proteins.

In this thesis the functionally diverse MHP multigene family of enzymes (MPO, EPO, LPO and TPO) were used as a case study to better understand the evolution of specificity. This thesis tests two hypotheses involving phylogenetic reconstruction and enzyme evolution. Firstly, the hypothesis that positive selection and functional shift are synonymous is tested. By fully resolving the evolutionary relationship of the MHP from a molecular perspective, the patterns of gene duplication and subsequent selective forces that have contributed to the functional shift in each type of MHP may be elucidated. To determine if such selective pressures are truly responsible for functional diversification a cross-disciplinary approach can be taken to validate the *in silico* (evolutionary) predictions. *In vitro* site-directed mutagenesis can be implemented to verify if altering identified positively selected residues to a more ancestral state disrupts the cellular and biochemical properties of the extant enzymes. This approach was applied here using the MHP superfamily, MPO as an exemplary model to test the hypothesis that positive selection signifies a protein functional shift.

Secondly, the hypothesis that ancestral protein reconstruction can be used as a tool to resurrect ancient proteins that contain the desirable properties of extant proteins is tested. The plant heme peroxidases, which have been extensively exploited in the

biopharmaceutical and biotechnology sectors, have been used as a case study in this thesis to determine if evolutionary theory (paleomolecular biochemistry) can be used in the design of enzymes with desirable characteristics for industry. The evolutionary relationship of this family of enzymes has been resolved, thus, it is possible to reconstruct and resurrect an hypothetical common ancestor of these family members to characterise biochemically.

Chapter 2

Phylogenetic and Selective Pressure Analyses of the Mammalian Heme Peroxidases.

2.1 Introduction

Heme peroxidases are readily abundant enzymes that can be classified into two major families, namely the animal and non-animal peroxidases, that have arisen from two independent evolutionary events (Passardi et al. 2007a). The non-animal peroxidases include plant, bacterial, fungal and protist (Passardi et al. 2007a). The classical peroxidase cycle involves the reaction sequence from native enzyme through compound I, then compound II and finally back to native enzyme (Dunford 1999). An alternative and highly important pathway that MHP pass through, depending on substrate availability, is the halogenation cycle (Furtmüller et al. 2006). In the presence of H_2O_2 and a halide (especially iodide), MPO can catalyse a halogenation reaction that plays an important role in the antibacterial activity of leukocytes (Clark 2000). Animal peroxidases are a medically important group of enzymes implicated in many different diseases including asthma (Sanz et al. 1997), AD (Reynolds et al. 2000) and inflammatory vascular disease (Lau, Baldus 2006). From biochemical studies it is believed that the heme peroxidases for mammals arose following a number of gene duplication events (Sakamaki et al. 2000; Sakamaki, Ueda, Nagata 2002; Furtmüller et al. 2006).

Gene duplication provides the raw material for evolution of diversity and is believed to be the principal source of new genes (Ohno 1970). The process of gene duplication has a number of alternative outcomes, and remains a controversial issue. Gene duplicates may become functionally redundant (Nowak et al. 1997), or functionally divergent. There are a number of ways in which functional redundant duplicates can be preserved (Force et al. 1999; Chung et al. 2006). It has been proposed that the preservation of duplicates can be brought about by degenerative mutations in the regulatory elements of the duplicates, this is referred to as the Duplication-Degeneration-Complementation model (DDC) (Force et al. 1999). The DDC model does not allow a role for positive selection in the evolution of duplicates and is based solely on a neutral model with degenerate mutations and subsequent negative selection. Under this model duplicates are preserved as each

accumulates degenerate mutations, resulting in specific subfunctions that *in toto* ensure optimal fitness (Force et al. 1999).

An alternative mode of duplicate retention is positive selection. For example, in direct contrast to the predictions of the DDC model it has been shown for human and mouse that the number of retentions and losses of duplicates fits more consistently with a model incorporating positive selection (Shiu et al. 2006). Rapid divergence in gene expression profiles of duplicates following the duplication event results in expression profiles as diverse as those of singletons. An example of this is the functional redundancy of transcription factor inhibitors, Iκα and β, that have acquired different functions through divergence of gene expression rather than biochemical function (Cheng et al. 1998). Studies have indicated that for mammalian genomes neofunctionalisation, be it independent of -, or coupled with – subfunctionalisation, is the most common mode of evolution of gene duplicates (Hughes, Liberles 2007). These selective pressures following the process of gene duplication are key to the evolution of specificity of divergent multigene families, such as the MHP (Kimura, Ikeda-Saito 1988).

In those cases where having all duplicates is deleterious, dosage requirements may cause the partitioning of subfunctions to be favored by positive selection resulting from selective pressure for the fixation of nonfunctional or subfunctional alleles. The divergence of function may occur through neofunctionalisation (Hughes 1999), or, subfunctionalisation where the ancestral function is partitioned between the duplicates (Lynch, Force 2000) (for detail on current gene duplication models see (Roth et al. 2007)).

This study hypothesises that the selective pressures on MHP following gene duplication events will, (i) still be traceable in the extant sequences of these enzymes, and (ii), will have contributed to the functional diversity observed in these enzymes. A fully resolved phylogeny can provide a basis for such comparative genomic analysis of these heme peroxidases.

MHP have been classified into four main families based on their function; MPO, EPO, LPO and TPO. MPO, EPO and LPO function in antimicrobial and innate immune responses (Klebanoff 1970; Klebanoff 1999; Wang, Slungaard 2006), whereas, TPO plays a key role in thyroid hormone biosynthesis (Ruf, Carayon 2006), see Table 1.1. A study of the structure-function relationships of human heme peroxidases suggest that the evolution of TPO succeeded that of MPO, EPO and LPO, but that these families shared a common ancestor (Sakamaki et al. 2000; Sakamaki, Ueda, Nagata 2002; Furtmüller et al. 2006). MHP are present in various tissues and as such their peroxidase function varies depending on tissue of expression. There are both structural and functional similarities among this multigene family of enzymes particularly with respect to their catalytic domains, this reflects their evolutionary relatedness. It has been shown that active site residues are conserved in all heme peroxidases (Furtmüller et al. 2006; Zederbauer et al. 2007).

To infer the phylogeny of the MHP from sequence data, it is fundamental to consider the challenges associated with resolving mammalian gene phylogenies. The main pitfalls include poor phylogenetic signal resulting from mutationally saturated positions, inadequate modelling of the evolutionary process and systematic bias due to variable rates of evolution among species or within sequences (Moreira, Philippe 2000).

A systematic bias or systematic error is one that results in greater support for an incorrect conclusion with the accumulation of more data. LBA is one of the most commonly occurring systematic biases and is a consequence of unequal evolutionary rates across lineages. This can occur due to the number of cell divisions per unit time being different in different species or due to rapid fixation of mutations due to reduced population size, e.g., a bottleneck. Rodent species accumulate many more mutations within a defined time frame than larger mammals (Ohta 1993; Li et al. 1996). Therefore, rodentia are often placed close to the outgroup species on a phylogeny due to their increased number of mutations. There are a number of ways in which the noise (LBA) can be minimised. Firstly, the addition of more taxa to the dataset: denser sampling of species of intermediate generation time can reduce the effect of LBA by reducing the

overall distances between taxa. Secondly, the use of improved models of sequence evolution, i.e., models sensitive to multiple substitutions at the same site and rate heterogeneity across the phylogeny. And finally, stripping the alignment of its most rapidly evolving sites and using only the remaining more slowly evolving sites to reconstruct phylogenies reduces the amount of LBA noise in the dataset (Philip, Creevey, McInerney 2005). These approaches can be used in combination. While databases such as Peroxibase (http://peroxibase.isb-sib.ch) house all the up-to-date peroxidase sequences (Bakalovic et al. 2006; Passardi et al. 2007b; Koua et al. 2009; Oliva et al. 2009), only those MHP from completed mammalian genomes have been included here (allows us identify species-specific gene birth and death). ML and Bayesian methods of phylogeny reconstruction have been implemented together with the stripping of the most rapidly evolving sites in the dataset.

The major questions addressed in this study pertain firstly to the resolution of the evolutionary relationships of these MHP using molecular sequence data, and secondly, to the analysis of functional diversities among these superfamilies using the resolved phylogeny and ML methods for testing selective pressures.

Selection can be classified as being neutral, purifying or positive. Positive selection/Adaptive evolution is strongly indicative of functional shifts within proteins (Levasseur et al. 2006a). To determine what selective pressures may have influenced the functional diversification of the MHP families, we tested the data using a variety of ML models of evolution with different properties. These included models that allow for only purifying selection and/or neutral evolution, and those that allow for positive selection. Likelihood scores for all alternative models and their null hypotheses are calculated. The likelihood scores for the null hypothesis versus the alternative hypothesis for those models that are extensions of each other were then compared using a LRT for goodness-of-fit. For those models that allow for the estimation of site-specific evolution, those amino acids that have undergone positive selection can be identified. The location of these amino acid positions were estimated using Bayesian statistics and their location and possible functional significance were determined. This chapter has shown that

positive selection has contributed to the evolution of these enzymes following gene duplication events.

2.2 Methodology

2.2.1 Data Assembly

2.2.1.1 Sequence Data

Protein coding sequences for MHP were retrieved from the Ensembl database for all available completed mammalian genomes using the pre-defined orthologues identified in Ensembl (www.ensembl.org). The mammalian genomes and the corresponding genome versions used for each of the major families in our dataset were as follows: *Homo sapiens* v42.36d; *Pan troglodytes* v42.21a; *Macaca mulatta* v42.10b; *Mus musculus* v42.36c; *Rattus norvegicus* v42.34l; *Canis familiaris* v42.2; *Bos taurus* v42.2e (no EPO sequence available), and, *Monodelphis domestica* v42.36c. Ensembl identifies orthologues by performing a genome-wide reciprocal WUBlastp+SmithWaterman search of each gene across all completed genomes. MSA is then performed using the MUSCLE software (Edgar 2004) and the best reciprocal hits following the sequence similarity search. The longest alternative transcript in each case was used. These sequences were combined into a single MHP dataset of 31 sequences. Two amino acid sequences representing the peroxidasin (PXDN) family, from the *Pan troglodytes* and the *Gallus gallus* genomes, were retrieved from the PeroxiBase database (http://peroxibase.isb-sib.ch). The sequence data are given in Table 2.1.

2.2.1.2 Multiple Sequence Alignment

Each protein coding sequence in the MHP dataset was translated to amino acid using in-house translation software. This protein sequence dataset and the two PXDN sequences were combined to give a dataset of 33 sequences (complete dataset). Both MHP and "complete" datasets were aligned in ClustalW 1.8 (Thompson, Higgins, Gibson 1994) independently using default parameter settings. The corresponding nucleotide sequences for the MHP dataset were aligned with respect to the amino acid MSA with the use of in-house software to insert gaps in the protein coding sequence according to their positions in the amino acid alignment. The nucleotide and subsequent protein MSAs were manually edited by removing ambiguous regions from the alignment using

Table 2.1: Representative mammalian heme peroxidase sequences used in this study.

Superfamily	Species	Entry ID (Name)*/Gene ID	Length (aa)
MPO	*Homo sapiens*	ENSG00000005381	778
	Pan troglodytes	ENSPTRG00000009449	778
	Macaca mulatta	ENSMMUG00000002266	777
	Mus musculus	ENSMUSG00000009350	719
	Rattus norvegicus	ENSRNOG00000008310	719
	Canis familiaris	ENSCAFG00000017474	743
	Bos taurus	ENSBTAG00000012783	596
	Monodelphis domestica	ENSMODG00000014737	403
EPO	*Homo sapiens*	ENSG00000121053	716
	Pan troglodytes	ENSPTRG00000009446	716
	Macaca mulatta	ENSMMUG00000011973	717
	Mus musculus	ENSMUSG00000052234	717
	Rattus norvegicus	ENSRNOG00000008707	716
	Canis familiaris	ENSCAFG00000017456	752
	Monodelphis domestica	ENSMODG00000014755	725
LPO	*Homo sapiens*	ENSG00000167419	713
	Pan troglodytes	ENSPTRG00000009448	712
	Macaca mulatta	ENSMMUG00000002264	716
	Mus musculus	ENSMUSG00000009356	711
	Rattus norvegicus	ENSRNOG00000008422	710
	Canis familiaris	ENSCAFG00000024533	719
	Bos taurus	ENSBTAG00000012780	713
	Monodelphis domestica	ENSMODG00000014744	719
TPO	*Homo sapiens*	ENSG00000115705	934
	Pan troglodytes	ENSPTRG00000011610	857
	Macaca mulatta	ENSMMUG00000009662	839
	Mus musculus	ENSMUSG00000020673	915
	Rattus norvegicus	ENSRNOG00000004646	915
	Canis familiaris	ENSCAFG00000003217	932
	Bos taurus	ENSBTAG00000002567	869
	Monodelphis domestica	ENSMODG00000014296	872
PXDN	*Pan troglodytes*	5828 (PtroPxd01)*	1463
	Gallus gallus	4049 (GgaPxd01)*	1447

Note: The common names for the genomes used are; *Homo sapiens*: human, *Pan troglodytes*: chimp, *Macaca mulatta*: macaque, *Mus musculus*: mouse, *Rattus norvegicus*: rat, *Canis familiaris*: dog, *Bos taurus*: cow, *Monodelphis domestica*: opossum, *Gallus gallus*: chicken. aa: amino acid. *Note: * - Assigned entry ID and Name in the PeroxiBase database* (http://peroxibase.isb-sib.ch).

the sequence alignment editor, Se-Al 2.0a11 (Rambaut 1996). The PXDN sequences served as an outgroup for the MHP and therefore aided in determining the earliest diverging MHP.

2.2.2 Phylogeny Reconstruction

2.2.2.1 Site Stripping and Phylogeny Reconstruction

The phylogenetic tree for the dataset was estimated using Bayesian statistics implemented in MrBayes 3.1.2 (Ronquist, Huelsenbeck 2003). The model of amino acid substitution used was JTT (Jones, Taylor, Thornton 1992) because following model testing using MultiPhyl (Keane, Naughton, McInerney 2007) this was the model that was best-fit to the data. Using 4 Markov chains for 400,000 generations, trees were sampled every 10 generations with the first 20,000 sampled trees discarded as burnin (the burnin parameter determines the number of samples (not generations) discarded prior to calculating summary statistics) (see Appendix for Bayes block template). The remaining trees sampled were summarized on a majority rule consensus tree with clade supports given as Posterior Probabilities (PPs). ML trees were also inferred using the high-throughput phylogenomics webserver, MultiPhyl (Keane, Naughton, McInerney 2007). The ML tree was generated using the nearest neighbour interchange (NNI) tree search algorithm and 100 bootstrap replicates implemented in MultiPhyl (Keane, Naughton, McInerney 2007) under the AIC statistic, the selected substitution model was JTT with invariable sites and a discrete gamma model of rate heterogeneity. This was repeated a total of 10 times to generate 1000 bootstrap replicates. (The Bayesian tree reconstruction methods were applied to the MHP dataset only).

The resulting phylogenies from both analyses (MrBayes and MultiPhyl) were then analysed for signatures of LBA. The rate of evolution at each site in the alignment was placed into one of 8 categories, 8 being the most rapidly evolving and 1 being the most conserved, using the maximum likelihood approach implemented in TreePuzzle 5.1 (Schmidt et al. 2002). Sites were progressively removed from the protein MSA

according to their evolutionary rate and the resultant trees were analysed for changes in topology.

Nine separate site-stripped alignments were constructed by successive removal of the most rapidly evolving sites (Philip, Creevey, McInerney 2005). The aforementioned Bayesian method was used to infer phylogenetic relationships for each of the nine alignments generated. The ML phylogeny was also estimated for each of the site-stripped alignments from the model of best-fit following hierarchical likelihood ratio tests (hLRTs) of alternative models implemented in MultiPhyl (Keane, Naughton, McInerney 2007).

2.2.2.2 Nodal Distance Analysis

The pruned nodal distance method implemented in TOPD/FMTS v3.3 (Puigbo, Garcia-Vallve, McInerney 2007) was used to calculate the distance between each of the site-stripped trees and the ideal tree. The ideal tree was generated by pruning the resolved mammalian phylogeny (Murphy et al. 2001) to represent those taxa present. A distance matrix is calculated for both the site-stripped phylogeny and the ideal phylogeny by counting the number of nodes that separate every taxon from every other taxon on the tree. Using the root means squared deviation (RMSD) implemented in the TOPD/FMTS v3.3 (Puigbo, Garcia-Vallve, McInerney 2007) software package, the RMSD between the site-stripped phylogeny matrix and the ideal phylogeny matrix is calculated. A RMSD value of zero indicates that the two trees being compared are identical (Input file in Appendix).

2.2.2.3 Gene Tree - Species Tree Reconciliation

Following nodal distance analysis, the gene phylogeny with the lowest RMSD value (for the MHP sequences alone), and the species tree were examined for gene duplication and loss events using the default settings for gene tree - species tree reconciliation implemented in GeneTree 1.3.0 (Page 1998).

2.2.3 Positive Selection and Functional Divergence

2.2.3.1 Selective Pressure Analysis

Analysis of variation in selective pressure following gene duplication in the MHP was carried out using codon substitution models implemented in PAML 3.15 (Yang 1997). Both site-specific and branch-site specific models were applied. The models used for this analysis allow for heterogeneous nonsynonymous-to-synonymous rate ratios ($\omega = Dn/Ds$) across sites and amongst branches/lineages.

An ω-value > 1 indicates positive selection, ω < 1, purifying selection and neutral evolution when ω = 1. The statistically significant model for the data was selected using a series of LRTs to compare models and their more parameter rich extensions. Tests of significance were carried out using χ^2 tests of significance, the comparisons performed were; M0 (one ratio) with M3(k = 2)(discrete), M1(neutral) with M2(selection), M3(k = 2) with M3(k = 3) discrete models, M7 (beta) with M8 (beta & omega > 1), M8 (beta & omega > 1) with the null hypothesis M8a (beta & omega = 1), M1 with model A (branch-site) and finally M3(k = 2) with model B (branch-site). The models and approach taken here have been described previously (Yang 1997; O'Connell, McInerney 2005).

The probability (PP) of a specific amino acid site belonging to the positively selected category is estimated using the empirical Bayes method for each superfamily individually (Nielsen, Yang 1998; Yang et al. 2000; Yang, Wong, Nielsen 2005).

2.2.3.2 Functional Divergence analysis

Using the MHP gene phylogeny with the lowest RMSD value, each of the four MHP were selected as independent clusters. Using the MHP protein MSA and this MHP gene phylogeny, statistical analysis implemented in the software DIVERGE v 1.04 (Gu 1999; Gu, Vander Velden 2002), was used to estimate the coefficient of functional divergence (theta ML or θ) for all pairs of clusters. The following are the clusters used in the

analysis are taken from the resolved phylogeny (from Figure 2.3a) (1) MPO Cluster, (2) EPO Cluster, (3) LPO Cluster, and (4) TPO Cluster.

2.2.3.3 3D Modeling and *In Silico* Mutational Analysis

To successfully build an homology model requires that a pre-existing 3D structure (template) has a high level of amino acid sequence similarity with the sequence of interest being modeled. Homology modeling was performed using the human representative sequence for the MPO, EPO and LPO family and the first approach mode implemented by the homology-modeling server, SWISS-MODEL (Arnold et al. 2006). A model for human TPO could not be achieved as a crystal structure of TPO has not yet been determined and there was minimal sequence coverage with other MHP template models. The human MPO and EPO structures were modeled using the crystal structure of bromide-bound human MPO isoform C (PDB accession code 1d2vC) as a template with 100 % and 72 % sequence identity to the template respectively and the human LPO structure was modeled using the crystal structure of bovine LPO (PDB accession code 2gj1) with 85 % sequence identity. The positively selected sites identified from the PAML 3.15 (Yang 1997) analysis were highlighted (in gold) on each 3D structure generated using DeepView v3.7 (Guex, Peitsch 1997). The conserved proximal heme ligand (MPO: His 502, EPO: His 474 and LPO: His 468) was also highlighted (in blue) on the 3D model.

In silico mutational analysis on these positively sites was carried out and their subsequent effect on hydrogen bonding was assessed using the mutate tool within DeepView application (Guex, Peitsch 1997). This application tool allows one to ascertain the putative effect of mutating the amino acid *in silico*. A nearest neighbour list of amino acid residues to the residue of interest were selected, with a cutoff of 8 Å as this is greater than the longest sidechain, tryptophan (Martin et al. 2002). Hydrogen bonds were computed between groups using the application tool provided. Using the mutate tool, the residue of interest was mutated to an alternative amino acid and putative hydrogen bonds were determined.

2.3 Results

2.3.1 Mammalian Heme Peroxidase Phylogeny

2.3.1.1 Phylogeny Reconstruction

The MHP dataset for this study consisted of 31 single gene orthologues from MPO, EPO, LPO, and TPO classes, totaling 1,017 aligned positions. The species phylogeny for the mammals has previously been fully resolved (Murphy et al. 2001). In brief, the mammalian species phylogeny describes Marsupiala (i.e. Opossum in our dataset) as outgroup to all other mammals, followed by the divergence of the Carnivora (i.e. Dog in our dataset) and the Cetacea (i.e. Cows in our dataset), and finally the emergence of the Euarchontoglires clade (i.e. primates and rodents) (Murphy et al. 2001), see Figure 1.6 and 2.1(a). The ML phylogenetic tree was estimated using MultiPhyl (Keane, Naughton, McInerney 2007) and MrBayes 3.1.2 (Ronquist, Huelsenbeck 2003), the results were congruent, see Figure 2.1(a). Each of the four superfamilies branched into their respective functional groups, with the members of the TPO superfamily taking the position of outgroup with high support values. The topology shows MPO, EPO and LPO shared a most recent common ancestor (MRCA) with a gene duplicate of TPO. The MPO and EPO groups themselves shared a MRCA and functionally diverged following a further gene duplication event. Therefore these two peroxidases (MPO & EPO) are the most closely related of all the MHP in this study.

2.3.1.2 Long Branch Attraction

Despite the 4 major clades in the phylogeny corresponding to the 4 major groups of MHP, the relationships of the species within these clades conflicts with the previously published mammalian species phylogeny (Murphy et al. 2001). The rat and mouse are members of the Glires group, and as such are a sister group to the primates, which together form the Euarchontoglires mammalian superorder. The topology seen here for the LPOs (see Figure 2.1(a)) suggests that dog and cow are the outgroup to the primate

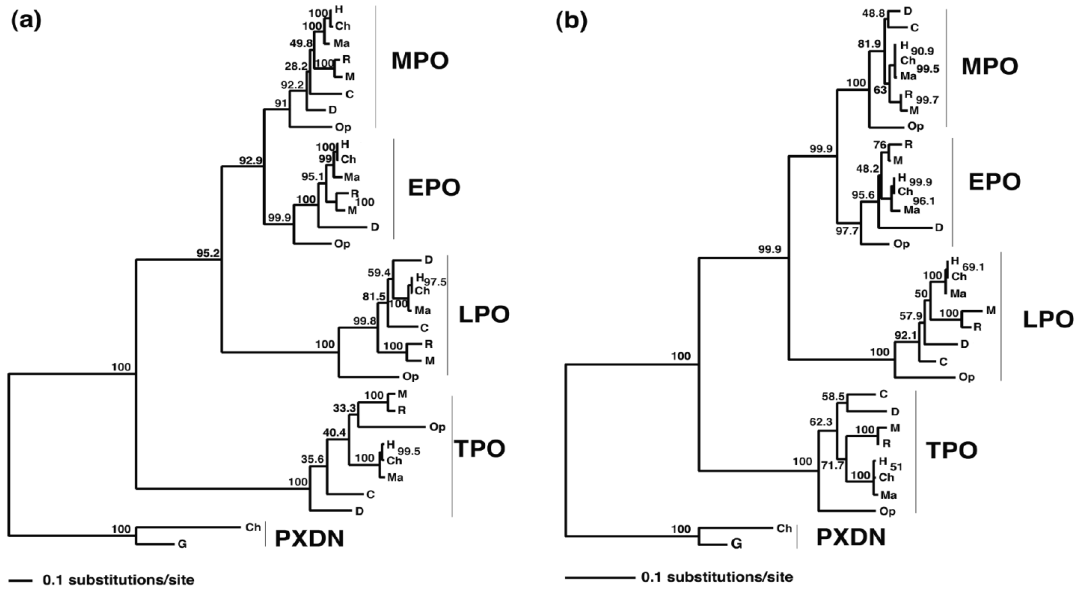


Figure 2.1. Phylogeny of mammalian heme peroxidases before treatment for long branch attraction and after treatment. *(a)* Initial unresolved ML tree for mammalian heme peroxidases and peroxidasin from *Pan troglodytes* and *Gallus gallus* from the entire dataset. The bootstrap support values from 1000 replicates are shown on all nodes. *(b)* Resolved phylogeny following site stripping, the cow sequence for LPO can be seen to take an unusual place on the phylogeny. The following are the species abbreviations used: Dog (D); Cow (C); Macaque (Ma); Human (H); Chimp (Ch); Rat (R); Mouse (M); Chicken (G), and Opossum (Op).

clade. This is a common error in mammalian phylogeny reconstruction, and has been proven to be an effect of LBA (Lunter 2007). Also, for the TPO group opossum is placed next to rat and mouse and not as the outgroup as expected, suggesting that the opossum and the rodents have similar rapid rates of evolution, see Figure 2.1(a).

The site stripping method was adapted by using the slow-evolving positions for each species in the MSA to reconstruct the phylogeny, while still retaining adequate amounts of signal (Philip, Creevey, McInerney 2005). This approach is similar to the '*Slow-Fast Method*' (Brinkmann, Philippe 1999) and is therefore an approximate method that removes noise from the data by removing those sites that are most likely to contain homoplasy (characteristics shared by a set of species but not derived from a common ancestor) and focusing on the more evolutionary informative positions for phylogeny reconstruction. Each site within the MSA was classified according to rates of evolution (estimated using ML based on a fixed phylogenetic tree). To determine what number of categories to remove, each category was progressively stripped from the most rapidly evolving sites to the most slowly across the entire MSA. The combined removal of the fastest and slowest sites from the dataset was also carried out, this was initially performed with the PXDN data included. Each time a category was removed the phylogenetic tree was estimated from the remaining MSA using ML.

2.3.1.3 Resolved Phylogeny

The ideal tree was created by pruning the mammalian supertree as published by Murphy *et al*. (2001) (with the inclusion of chicken) and is depicted in Figure 2.2(a). The difference between each site-stripped phylogeny and the ideal phylogeny was calculated using a nodal distance calculation RMSD (Puigbo, Garcia-Vallve, McInerney 2007), see Figure 2.2(b). From Figure 2.2(b), it is seen that the removal of rapidly evolving sites gradually removes the noise from the data and the remaining signal moves towards the canonical species phylogeny (Murphy et al. 2001). For the dataset consisting of MHP and PXDN sequences, the RMSD value reaches a minimum at the removal of 4 site

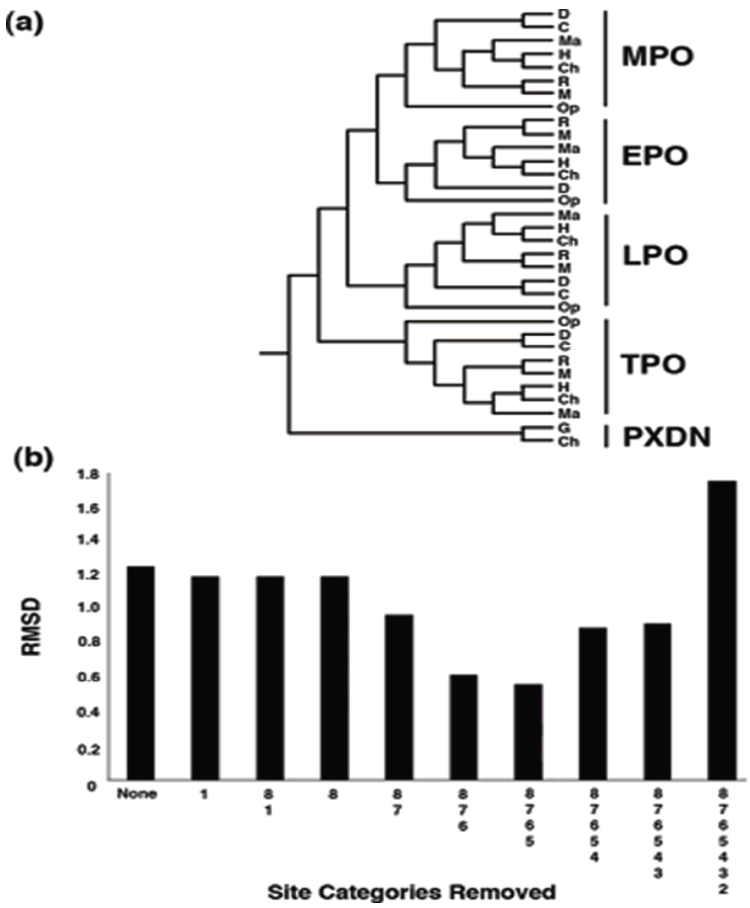


Figure 2.2. **The distance between each of the site stripped phylogenies and the ideal mammalian peroxidase phylogeny.** *(a)* The ideal phylogeny pruned from the mammalian phylogeny by Murphy *et al.* (2001), the peroxidasin sequences are outgroups to the MHP clade. The following are the species abbreviations used: Dog (D); Cow (C); Macaque (Ma); Human (H); Chimp (Ch); Rat (R); Mouse (M); Chicken (G), and Opossum (Op). This phylogeny was compared to each of the resultant site stripped phylogenies. *(b)* Graph showing the RMSD nodal distance (*y-axis*) between each site-stripped phylogeny (*x-axis*) and the ideal phylogeny. *On the X axis*: All: refers to the complete MSA; 8: site category 8 removed from the MSA; 8, 7: categories 8 and 7 removed from the MSA and so on up to the final column that contains only the most slowly evolving category of site.

categories (8, 7, 6 and 5) leaving a MSA of length 850 sites (including gaps/missing data), after this point the RMSD values rise, see Figure 2.2(b). It is important to note that the slowest evolving positions can be misleading particularly with excessive removal of sites, as the number of characters for reconstruction will decrease with every cycle, therefore caution must be taken in applying this method. This analysis was also performed on the dataset containing only MHP sequences, and the RMSD value reaches a minimum at the removal of 3 site categories (8,7, and 6) leaving a MSA of length 613sites (including gaps/missing data), see Figure 2.3(a) for resultant topology. The reduced MSA for MHP data is given in the Appendix additional Figure 1 and the corresponding TOPD results are given in Appendix additional Table 1. The nodal distance (RMSD) calculation is based entirely on the branching pattern and hence does not account for evolutionary rate variation across the phylogeny. Using this site-stripped MSA the phylogeny was estimated using both MrBayes and MultiPhyl methods, both of which produced identical phylogenies*. (*Of note here is that the one exception, using the Bayesian reconstruction method, was the TPO primate monophyly was not fully resolved in the TPO clade but instead supported a human-chimp-macaque polytomy.)

2.3.1.4 Gene Duplication

All gene duplication events were verified using gene tree – species tree reconciliation. The resolved MHP phylogeny was analysed (Figure 2.3(a)), identifying in total 4 duplication events and 4 losses. This method over prescribes gene losses as in the case of EPO, where the sequence data was not available and therefore is assumed to be a loss. There is an LPO specific duplication event predicted, see Figure 2.3(b). These results show differential retention and loss in the LPO lineage following this gene duplication event resulting in the cow species retaining an alternative duplicate copy to the other mammals in the dataset, as shown in Figure 2.3(b). This method must be used with caution as it does not take into account rate heterogeneity amongst species or sites in the data, and relies solely on the topology. However, reciprocal BLAST analysis of the cow sequence against the other mammal genomes identifies this sequence as an ortholog.

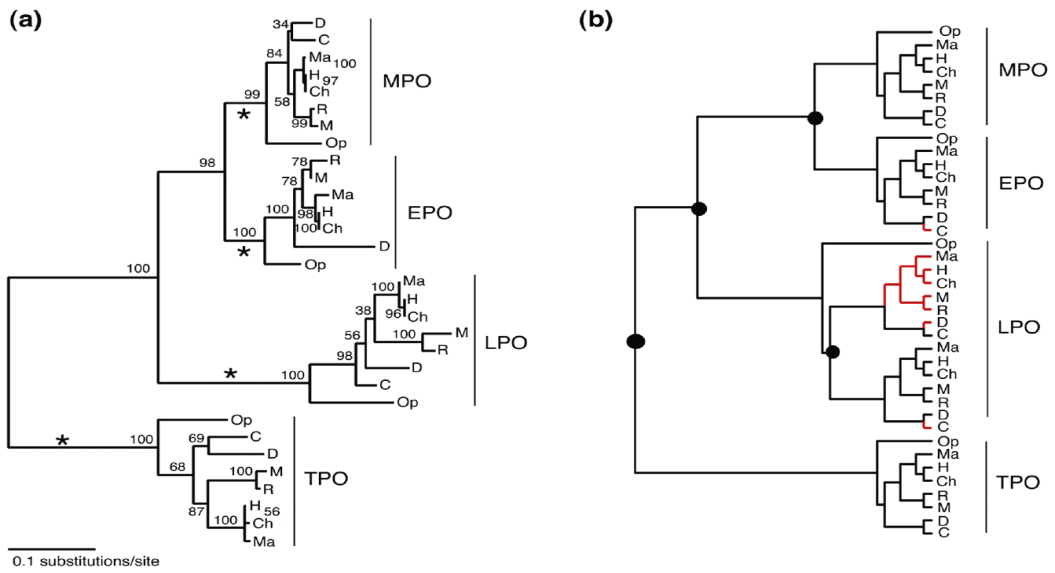


Figure 2.3. **Fully resolved mammalian heme peroxidase phylogeny with duplication and loss events depicted.** *(a)* Resolved ML tree for mammalian heme peroxidases. The bootstrap support values from 1000 replicates are shown on all nodes. The TPO primate clade appears here as a polytomy as the branch lengths are extremely short, however, this is in fact resolved with a low Bootstrap of 56%. The star symbol denotes those branches that were treated as foreground in the selection analysis. *(b)* The analysis of the resolved phylogeny using gene tree species tree reconciliation method implemented in GeneTree. The large filled circles represent gene duplication events, and the red branches indicate gene losses. The following are the species abbreviations used: Dog (D); Cow (C); Macaque (Ma); Human (H); Chimp (Ch); Rat (R); Mouse (M); Chicken (G), and Opossum (Op).

2.3.2 Positive Selection and Functional Divergence

2.3.2.1 Positive Selection Analysis

Using the fully resolved MHP phylogeny the hypothesis that following the gene duplication events in the MHP (as resolved in this study), selective forces - specifically positive selection - have contributed to the observed changes in function in each of the 4 major groups of MHP was tested. Tests for heterogeneous selective pressures were carried out on the resolved phylogeny using the evolutionary models implemented in PAML 3.15 (Yang 1997) and the complete MSA. The Dn/Ds ratios were estimated in a likelihood framework at both site-specific and lineage-specific levels. A total of seven tests of significance were carried out using χ^2 tests of significance, five site-specific comparisons and two branch-site comparisons were performed.

No positively selected sites were estimated for the one ratio model (see Table 2.2). Strong purifying selection across sites was indicated with an ω of 0.1516. However, this model is a poor fit for the data (ln*L* = -34417.1085). Positive selection was tested in a site-specific manner across the dataset using the site models; M1 (neutral), M2 (selection), M3 discrete (k = 2), M3 discrete (k = 3), M7 (beta), M8 (beta & omega > 1) and M8a (beta & omega = 1). The results of the site-specific analysis are shown in Table 2.2.

Poor likelihood values were achieved using the site-specific models of evolution, however, the most complex site-specific model used, M8 yielded significant results when it was tested with its null model M8a. A small proportion of sites are under relaxed positive selection (Table 2.2). Through the use of Bayesian estimations, four positively selected sites have been identified across the alignment, with posterior probability (PP) > 0.50.

Table 2.2: Parameter estimates and likelihood scores of one ratio and site-specific models.

Model	P	L	Estimates of parameters	Positively selected sites
M0 : one ratio	1	-34417.1085	$\omega = 0.1516$	None
Site-specific:				
M1:Neutral	1	-33999.1059	$p_0 = 0.7685$	Not allowed
M2:Selection	3	-33999.0008	$p_0 = 0.7685$, $p_1 = 0.2306$ ($p_2 = 0.0009$), $\omega_2 = 1.0000$	None
M3:Discrete(K = 2)	3	-33666.6464	$p_0 = 0.5205$, ($p_1 = 0.4795$) $\omega_0 = 0.0464$, $\omega_1 = 0.3311$	None
M3:Discrete(K = 3)	5	-33555.0100	$p_0 = 0.2272$, $p_1 = 0.4418$, $p_2 = 0.3310$ $\omega_0 = 0.0062$, $\omega_1 = 0.1124$, $\omega_2 = 0.4242$	None
M7: Beta	2	-33545.1089	$p = 0.5950$, $q = 2.3340$	Not allowed
M8: Beta&Omega > 1	4	-33540.9163	$p_0 = 0.9849$, $p = 0.6204$, $q = 2.6031$ ($p_1 = 0.0152$), $\omega = 2.0778$	BEB 4 > 0.50
M8a: Beta&Omega = 1	4	-33542.7233	$p_0 = 0.9733$, $p = 0.6366$, $q = 2.801$ ($p_1 = 0.0268$), $\omega = 1.0000$	Not allowed

Note: BEB: Bayes Empirical Bayes analysis

Results of the branch-site model B with each of the families individually labeled as foreground are shown here in Table 2.3; see Figure 2.3a for corresponding foreground branches. (Results for model A are given in the Appendix Tables 2 and 3). To determine whether there is rate heterogeneity along different branches in the phylogeny, models allowing for only site-specific evolution were compared with those allowing for branch-site specific evolution (i.e. M3 K = 2 with Model B and M1 with Model A). Following LRT analysis it was found that both models A and B were significant following χ^2 test with two degrees of freedom. The likelihood score from model B for each family had improved significantly from those obtained using model A, as a result, model B was determined as the best fit model in each case tested and these results are summarized in Table 2.3.

Positively selected sites identified with model B were estimated using NEB method (Yang et al. 2000). The results of which are discussed now in detail.

These results show that following gene duplication, each individual type of MHP has undergone positive selection in amino acid residues that are unique to that type of MHP, see Table 2.3. As positive selection is closely associated with functional shift, we postulate that these positively selected sites have significantly contributed to the evolution of the functional diversity of these MHP.

For the MPO superfamily, a total of 19 positively selected sites were identified (PP > 0.50). We have found functional information from the literature on 11 of these sites, these are now discussed: Position 80 (Arg) is located within the propeptide sequence and is under positive selection. Previous studies indicate that propeptide in MPO plays a key role in the processing and sorting of human MPO (Andersson et al. 1998). Position 568 is under positive selection and is next to the polymorphic site R569W. Mutations in position 569 have been shown to suppress posttranslational processing in MPO (Nauseef 1989). The 2 positions with strongest support, PP > 0.95, are separated by 8 amino acid residues on the MPO heavy chain, they are Asn496 and Leu504. These 2 positions along

Table 2.3: Parameter estimates and likelihood scores for branch-site model, model B.

Model	P	L	Estimates of parameters	Positively selected sites
MPO Model B	5	- 33655.0405	p_0= 0.4975, p_1= 0.4553, (p_2= 0.0246, p_3= 0.0225) *Background:* ω_0= 0.0458, ω_1= 0.3307, ω_2= 0.0458, ω_3= 0.3307 *Foreground:* ω_0= 0.0458, ω_1= 0.3307, ω_2= 251.6783, ω_3= 251.6783	**Foreground:** NEB 19 > 0.50 2 > 0.95 1 > 0.99
EPO Model B	5	- 33647.5634	p_0= 0.4967 p_1= 0.4469, (p_2= 0.0297, p_3= 0.0267) *Background:* ω_0= 0.0464, ω_1= 0.3322, ω_2= 0.0464, ω_3= 0.3322 *Foreground:* ω_0= 0.0464, ω_1= 0.3322, ω_2= 774.6323, ω_3= 774.6323	**Foreground:** NEB 28 > 0.50 6 > 0.95 4 > 0.99
LPO Model B	5	- 33627.3508	p_0= 0.4431, p_1= 0.3884, (p_2= 0.0898, p_3= 0.0787) *Background:* ω_0= 0.0470, ω_1= 0.3414, ω_2= 0.0470, ω_3= 0.3414 *Foreground:* ω_0= 0.0470, ω_1= 0.3414, ω_2= 82.8559, ω_3= 82.8559	**Foreground:** NEB 96 > 0.50 18 > 0.95 11 > 0.99
TPO Model B	5	- 33639.5793	p_0= 0.4358, p_1= 0.3690, (p_2= 0.1057, p_3= 0.0895) *Background:* ω_0= 0.0479, ω_1= 0.3468, ω_2= 0.0479, ω_3= 0.3468 *Foreground:* ω_0=0.0479, ω_1= 0.3468, ω_2= 999.0000, ω_3= 999.0000	**Foreground:** NEB 82 > 0.50 8 > 0.95

Model B allows each foreground lineage to be tested independently of all other lineages, hence the four clusters (MPO, EPO, LPO, TPO - each in turn treated as foreground), and estimates 5 parameters (P) in total. p_0,p_1, p_2 and p_3 are proportions of sites in the dataset with the corresponding ω value, i.e, ω_0, ω_1, ω_2 and ω_3 for the foreground and the background lineages independently. The final column gives the estimated number of sites with posterior probabilities of greater than 0.50 of belonging to the positively selected category. *Note: NEB: Naïve Empirical Bayes analysis.*

with Tyr500 are in close proximity to the proximal heme ligand in MPO, His502 (Furtmüller et al. 2006). Position 259 (Leu) is located between two important distal residues, Gln257 and His261, involved in the formation of hydrogen bonds (Furtmüller et al. 2006). His261 has an important role in the formation of compound I, a redox intermediate of the peroxidase cycle (Dunford 1999). A further four sites (Leu630, Gln633, Glu652; (primates Lys652) and Asn654 (primates Lys654) were identified as positively selected, PP > 0.70, these are located within a disulfide bond linking helices 19 and 22 on the MPO heavy chain. Disulfide bonds are associated with the folding and stability of proteins and as such are significant to the overall function of that protein (Rietsch, Beckwith 1998).

For the EPO clade, 28 sites are positively selected, PP > 0.50. Functional information for 15 of these sites has been found. One of these, Asp71, is located in the EPO propeptide. The inferred phylogeny, shown in Figure 2.1(b), suggests that MPO and EPO are closely related enzymes, therefore it may be possible that the EPO propeptide may also be crucial for the function of EPO. The region separating the catalytic residues Arg377 and His474 (Furtmüller et al. 2006), contains 8 positively selected sites (PP>0.50). Arg377 is the conserved prominent distal amino acid associated with hydrogen bond formation. The proximal heme ligands His474 (EPO), His502 (MPO) and His468 (LPO), are conserved in all the MHP (Furtmüller et al. 2006; Zederbauer et al. 2007). Six of the 28 positively selected sites, Arg584, Gln588, Arg591, Ala618, Gly626 and Ala627, are located on the EPO heavy chain within a single disulfide bond region, this would suggest that they are structurally and functionally important to EPO. Position 441 has been identified as positively selected, this residue has also been noted as being polymorphic in the human population (Lys/Thr).

There are 18 positively selected sites for the LPO group (PP > 0.95). Functional information on 13 of these sites has been found. Residues Glu72, Asn87 and Trp91 are found in the LPO propeptide sequence and have a probability of greater than 0.95 of being positively selected. Residues Asn255, Phe282, Ser312, Ser352 and Glu355 are all located in the disulfide bond region (PP > 0.95). From biochemical analysis both

Arg372 (Arg377 in EPO) and His468 are believed to have catalytic properties, and are conserved in the MHP (Furtmüller et al. 2006; Zederbauer et al. 2007). Positive selection has been detected in His376 (PP > 0.99) just four amino acids downstream of the first of these catalytic residues (Arg372), interestingly this site is specific to the primate lineage. Also we have detected positive selection in Glu470 (PP > 0.98) adjacent to the second catalytic site (His468). Positive selection has also been detected in Asp700 which is a known genetic variant and Glu240 and Gln245 that are located to the right and left of a known human polymorphism A244T.

With the TPO clade treated as foreground, 8 sites are positively selected, PP > 0.95. Of these 8 sites, 6 are missing in the alternatively spliced TPO isoform 5, which exhibits incorrect protein folding (Ferrand, Le Fourn, Franc 2003). Asp228 (PP > 0.95), Ala232 and Ala242 (both PP > 0.50) are in the region of the TPO active site His239. Glu378 has also been identified as a novel mutational site (E378K) associated with the common inherited deficiency total iodide organification defect (TIOD) and is under positive selection in our analysis (Tajima, Tsubaki, Fujieda 2005).

2.3.2.2 Functional Divergence

Independent analysis for positive selection using DIVERGE software further supports our findings, see Table 2.4 for summary of results. Values greater than zero for the coefficient of functional divergence, θ, indicate a functional shift between clusters. Rate heterogeneity among sites varies with respect to the gamma distribution (α). We estimated θ for each of the four MHP clusters. This analysis shows significant functional constraints among the four MHP clades, with the null hypothesis $\theta = 0$ being rejected for all clusters analysed. The analysis of closely related MPO and EPO clusters result in the lowest θ value (0.2833 +/-0.0837), and both have microbicidal activity (Table 1.1). θ increases at least 1.5 fold for the more distantly related/functionally divergent clusters. These results provide statistical evidence of the diverse functions of these MHP enzymes.

Table 2.4: Summary of results of analysis using DIVERGE software.

	MPO/EPO	MPO/LPO	MPO/TPO	EPO/LPO	EPO/TPO	LPO/TPO
θ ML	0.2832	0.4504	0.4984	0.4552	0.4304	0.4280
SE θ	0.0837	0.0744	0.0783	0.1021	0.0950	0.0756
LRT θ	11.4512	36.6860	40.4815	19.8713	20.5223	32.0448
α ML	0.3034	0.4221	0.4172	0.4863	0.4654	0.5413

Each cluster analysed is shown in the columns of the table. **θ ML:** Coefficient of functional divergence. **SE θ:** Standard error of the estimate Theta. **LRT θ:** 2 log-likelihood-ratio against the null hypothesis of θ = 0. **α ML:** Gamma shape parameter for rate variation among sites.

2.3.2.3 3D Modelling and *In Silico* Mutational Analysis

The relationship between positive/directional selection and functional shift was further tested by analyzing the effect of mutating the residues unique to MPO, EPO and LPO to a more ancestral state (majority rule ancestral state) on their respective 3D structure. This was achieved by modeling the human protein sequence for each enzyme using SwissModel and using the mutate tool in DeepView v3.7, *in silico* site directed mutagenesis has been performed on those positively selected sites (Guex, Peitsch 1997; Arnold et al. 2006). No homology model was generated for the human TPO representative as a crystal structure for TPO is not yet available and sequence identity to other MHP models was minimal. We find that mutating these positions from their positively selected state to the majority rule ancestral state causes a variety of effects on the hydrogen bond formation within the 3D structure, see Table 2.5, 2.6 and 2.7 for a summary of the effects on hydrogen bonds in MPO, EPO and LPO respectively. Hydrogen bonds play an important role in maintaining the structural integrity of a protein, any disruption of such forces is likely to upset the balance between the structural and functional dynamics (Martin et al. 2002). Other non-covalent interactions (e.g. hydrophobic and van der Waals forces) and steric hindrances (not assessed) may be disrupted as a consequence of these mutations and, in turn, may further compromise the structural integrity of the enzyme of interest.

The structure for MPO with positively selected sites and the heme binding site (His 502) is shown in Figure 2.4(a). On mutating each of these 19 (PP > 0.50) positively selected amino acids we find that 4 bonds are lost and 4 are independently gained in the protein, for summary see Table 2.5. For the mutations: N496F, Y500F, and L504T, the positions of the losses and gains of hydrogen bonds are significant as these amino acid are in close proximity to the proximal heme ligand His502, shown in Figure 2.4(a). The mutation from leucine to threonine at position 504 results in the formation of an additional hydrogen bond between Gly501 and Leu504. Gly501 is directly bound to the proximal heme ligand. In addition, the N496F mutation illustrated in Figure 2.4(b), results in the loss of the hydrogen bond with Asn587. The Asn587 and His502 are connected by a

Table 2.5: Summary of results from SwissModel/DeepView analysis of MPO specific positively selected sites.

Mutation	Posterior Probability	Effect on Hydrogen Bond
C316S	0.815	-/+
S414A	0.600	-
A471R	0.738	+
P477G	0.948	=
N496F	0.999	-
Y500F	0.731	-
L504T	0.970	+
R529E	0.657	+
I568L	0.686	=
P584A	0.949	=
L630F	0.767	=
Q633L	0.737	=
L652V	0.840	=
L654G	0.921	=
S687T	0.648	+

Note: Mutation from positively selected site in MPO (using human model) to the amino acid present in TPO at that position (in cases where there was conflict the majority rule consensus at that position was taken). Posterior Probability values extracted using NEB analysis in model B Codeml. Effect on H-Bonds is classified as "+" if an increase in the number of bonds with positively selected amino acid, "-" if a hydrogen bond or a number of hydrogen bonds were lost with the positively selected site, and "=" refers to no affect on the hydrogen bond with the positively selected site.

Table 2.6: Summary of results from SwissModel/DeepView analysis of EPO specific positively selected sites.

Mutation	Posterior Probability	Effect on Hydrogen Bond
A288S	0.854	=
T381V	0.819	=
F434Y	0.776	=
K441P	0.649	=
R443A	0.998	-
A444F	0.769	=
T447Y	0.859	=
G449P	0.997	=
C455D	0.991	+
S456P	0.895	=
N457T	0.977	+
V462I	0.582	=
L468A	0.958	=
Y511K	0.686	=
A534Q	0.577	=
R584E	0.777	+
Q588E	0.550	-
R591T	0.928	+
A618G	0.713	=
G626R	0.646	=
A627G	0.503	=
A644L	0.994	=
R661A	0.552	-

Note: Mutation from positively selected site in EPO (using human model) to the amino acid present in TPO at that position (in cases where there was conflict the majority rule consensus at that position was taken). Posterior Probability values extracted using NEB analysis in model B Codeml. Effect on H-Bonds is classified as "+" if an increase in the number of bonds with positively selected amino acid, "-" if a hydrogen bond or a number of hydrogen bonds were lost with the positively selected site, and "=" refers to no affect on the hydrogen bond with the positively selected site.

Table 2.7: Summary of results from SwissModel/DeepView analysis of LPO specific positively selected sites.

Mutation	Posterior Probability	Effect on Hydrogen Bond
E240F	0.993	=
Q245D	0.992	=
N255P	0.997	=
F282A	0.993	=
S312P	0.999	=
S352A	0.970	=
E355P	0.999	=
H376V	0.991	=
E470T	0.988	+
E488H	0.974	+
H546T	0.985	+
S697D	0.978	+
D700T	0.999	+
A708R	0.977	+

Note: Mutation from positively selected site in LPO (using human model) to the amino acid present in TPO at that position (in cases where there was conflict the majority rule consensus at that position was taken). Posterior Probability values extracted using NEB analysis in model B Codeml. Effect on H-Bonds is classified as "+" if an increase in the number of bonds with positively selected amino acid, "-" if a hydrogen bond or a number of hydrogen bonds were lost with the positively selected site, and "=" refers to no affect on the hydrogen bond with the positively selected site.

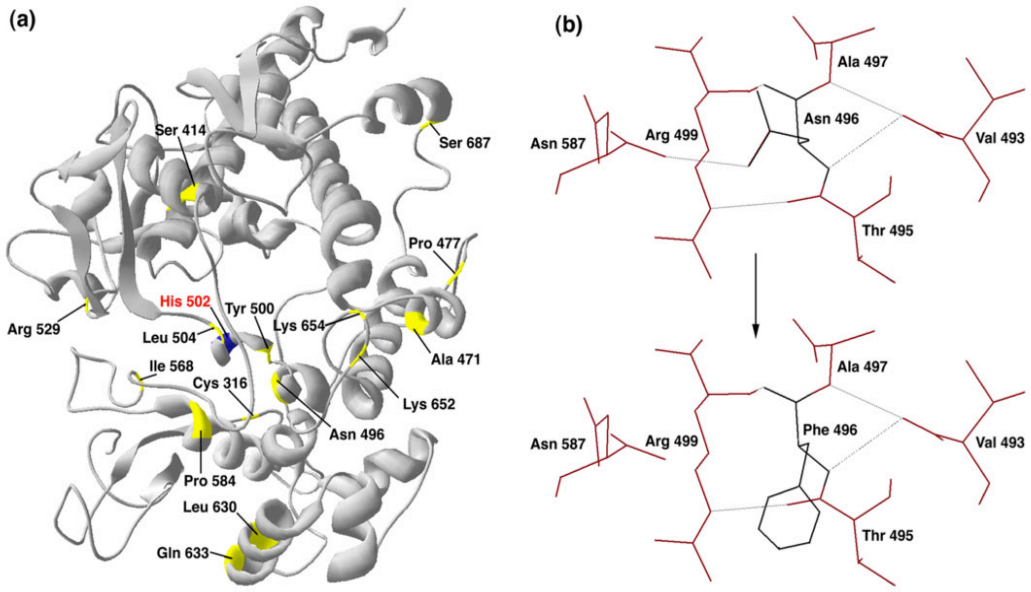


Figure 2.4. Location of positively selected sites in the myeloperoxidase structure and their effect on bonding within the structure. (a) 3-D structure of the human MPO sequence, highlighted in gold are those sites that are positively selected in MPO, in blue is the heme binding site. (b) Example of the effect on hydrogen bonding of one such mutation at positively selected position 496 in human MPO from Asparagine to Phenylalanine.

hydrogen bond (Furtmüller et al. 2006). The loss of the hydrogen bond, as a result of the mutation at position 496, is likely to affect the structural integrity of the link between Asn587 and His502. Disruption to the hydrogen bonds in this catalytically important region may have direct implications for functional divergence of the MPO enzyme. The A471R mutation results in an increase in the number of hydrogen bonds associated with this position. This position is upstream from Asn483 which is thought to be responsible for MPOs dimer interaction (Furtmüller et al. 2006). The mutation from cysteine to serine at position 316 results in the formation of a hydrogen bond with Gln329 and the loss of one of the bonds to Asp593, see Table 2.5. Cys316 is next to the single disulphide bridge (Cys319) that connects MPOs symmetry-related halves (Furtmüller et al. 2006). The C316S mutation may potentially disrupt this disulphide bridge.

For the EPO superfamily, 23/28 sites under positive selection with a PP > 0.50 were modeled due to the sequence coverage (72 %) between the human EPO sequence and the template MPO model selected (PDB accession code 1d2vC). See Figure 2.5 for the homology model of human EPO with positively selected sites and the heme binding site (His 474) shown. Mutating these residues to that present in its more ancestral counterpart, TPO, we find that 70 % of hydrogen bonds are uneffected, 3 bonds are disrupted and 4 additional putative bonds are formed, see Table 2.6 for a summary of the effects. Position Arg 443 has the highest probability of being under positive selection (PP = 0.998), and is within the region separating the two catalytically important residues R377 and H474. By mutating to alanine the hydrogen bond with Pro 631 is lost. There is a disulphide bond between C578 and C635, so the loss of the hydrogen bond between R443 and P631 may disrupt this disulphide bridge. Positions C455 and N457 have also a PP > 0.95, mutating these to aspartic acid and threonine respectively result in the formation of additional hydrogen bonds, potentially altering the enzyme's structure. The R661A mutation results in the loss of a hydrogen bond with position 655, which is adjacent to a know point mutation, D648N, that is associated with EPO deficiency (Nakagawa et al. 2001).

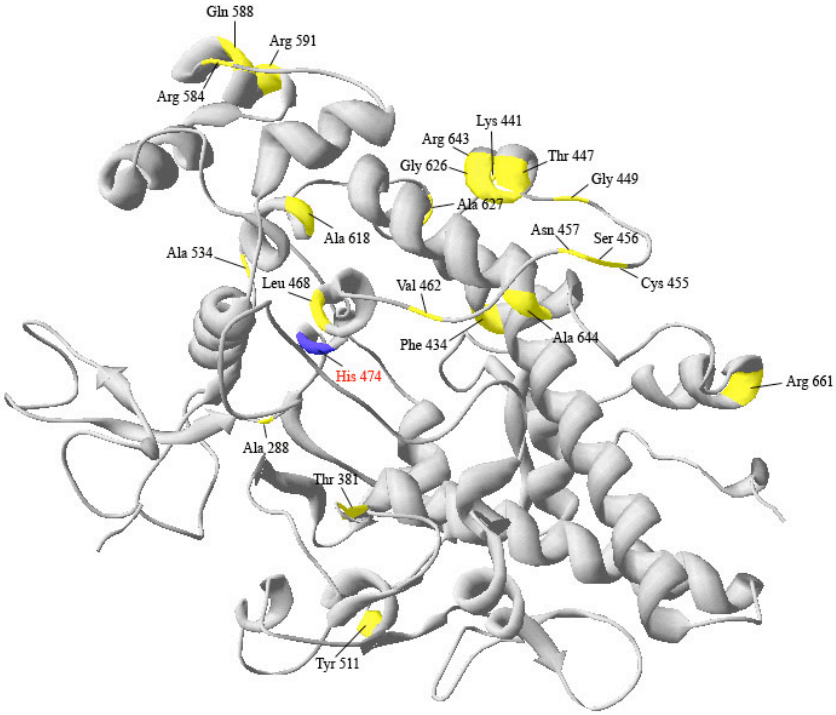


Figure 2.5. Location of positively selected sites in the eosinophil peroxidase structure. 3-D structure of the human EPO sequence, highlighted in gold are those sites that are positively selected in EPO, in blue is the heme binding site.

We identified a total of 96 amino acid residues under positive selection in the LPO superfamily (PP > 0.50). Due to this high number of sites we restricted our *in silico* mutational analysis to those sites with a PP > 0.95. These positively selected sites andthe heme binding (His 468) site are shown in Figure 2.6. The template selected to model the human LPO representative sequence was the crystal structure of LPO from cow (PDB accession code 2gj1) with 85 % sequence identity. On mutating these residues it is suggested that 6 independent hydrogen bonds are formed. When mutated to threonine, position 470 forms a hydrogen bond with position 467. Both these positions are neighbouring the proximal heme ligand, His 468. The mutation H546T results in the formation of a hydrogen with Lys 537. H546T results in an increase in hydrogen bonds associated with this position. His 546 is 8 residues from Asn 554, which forms a hydrogen bond with heme ligand His 468 (Furtmüller et al. 2006). A disulphide bridge is formed between positions 671 and 696. The positively selected site S697, D700 and A708 are in close proximity to this bridge and when mutated independently, each result in the formation of hydrogen bonds. Residues 695 and 697 are connected via a hydrogen bond, therefore, mutating serine to aspartic acid at position 697 may disrupt the aforementioned disulphide bridge.

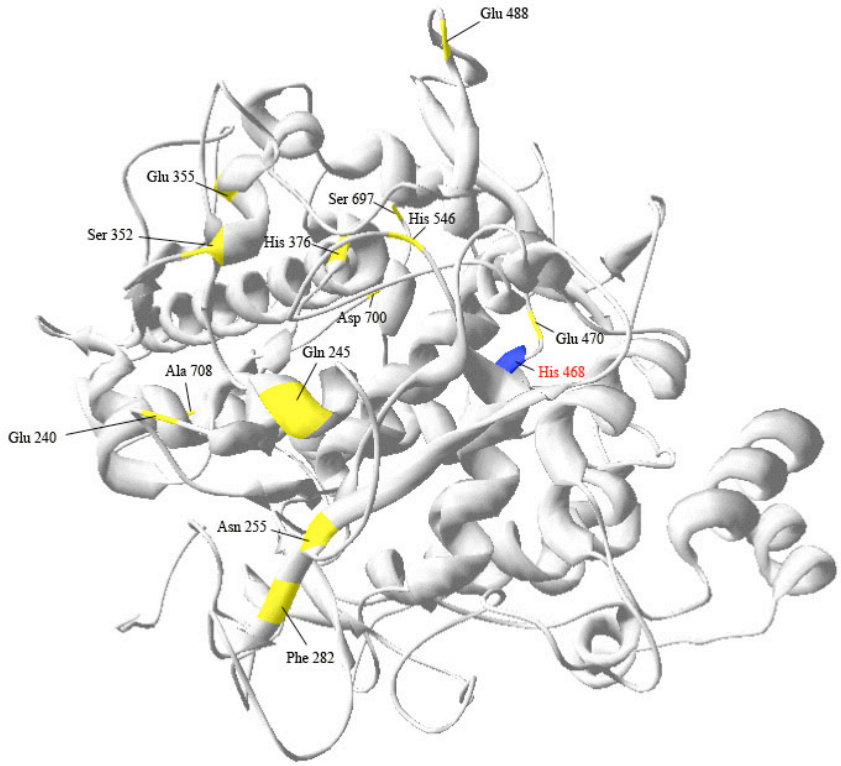


Figure 2.6. Location of positively selected sites in the lactoperoxidase structure. 3-D structure of the human LPO sequence, highlighted in gold are those sites that are positively selected in LPO, in blue is the heme binding site.

2.4 Discussion

The MHP are a functionally diverse family of enzymes which are implicated in a variety of inflammatory and neurodegenerative diseases such as asthma and AD respectively. In this study the evolutionary history of the four major groups of MHP; MPO, EPO, LPO and TPO, was investigated allowing for the analysis of their functional diversity.

Initial ML and Bayesian phylogenies estimated here for the MHP support previous biochemical studies (Sakamaki et al. 2000; Sakamaki, Ueda, Nagata 2002; Furtmüller et al. 2006). From Figure 2.3 the order of gene duplication events can be traced, with an MPO-EPO-LPO MRCA arising from a gene duplication with extant TPO; then a further duplication event that gave rise to, (i) the MPO-EPO MRCA, and (ii), the lineage leading to extant LPO; and the final and most recent duplication of the MPO-EPO MRCA into extant MPO and EPO clades. Peroxidasin is the outgroup to the MHP sequences and was included in the analysis to illustrate that TPO is the most ancestral MHP (Figure 2.1a). However, the species relationships estimated within these clearly defined clades were in disagreement with the previously resolved mammalian phylogeny (Murphy et al. 2001).

Including all sites of the alignment in the analysis, it is evident that the major types of MHP form monophyletic clades and are therefore the result of gene duplication events prior to speciation of modern day mammals; see Figure 2.1(a). However, also evident from Figure 2.1(a), species with more similar generation times are clustered together, with species of shorter generation times and therefore more rapid rates of mutation assuming a basal position in the phylogeny. This observed branching pattern could be a result of LBA, incorrect ortholog prediction or hidden paralogy.

If a phylogeny is seen to approach the ideal by removing the most rapidly evolving sites, then it is proposed that LBA is most likely to have contributed to the misleading phylogeny. To test for the presence of LBA, 8 categories of rates of evolution for all sites were calculated, from the most rapidly evolving to the most slowly evolving. It was

observed that the sequential removal of rapidly evolving categories of sites from the alignment decreased the difference, in terms of nodal distance RMSD, between the phylogeny produced and the ideal phylogeny. This occurred only for removal of the 4 fastest evolving categories of site from the alignment. Further removal after this point resulted in increased RMSD values between the phylogeny produced and the ideal. The MHP phylogeny shown in Figure 2.3(a), with maximum number of sites and minimum amount of noise. It is proposed that a possible reason for the presence of LBA in this dataset is the presence of taxa with vastly different generation times. The rodentia have previously been shown as "fast evolving" due to their short germ-line generation time, whereas species such as dogs and humans have longer germ-line generation times (Li, Tanimura, Sharp 1987; Ohta 1993; Li et al. 1996). In any given dataset there are sites that are variable and sites that are invariable, this pattern is conserved across homologous sequences. In a dataset with a mixture of germ line generation times, the mutation rate in the species with shorter germ line generation times will be higher, because the number of cell divisions per unit time is greater. Therefore the number of mutations in the variable regions will increase for these species. The result is an LBA effect derived from having a mixture of long and short germ line generation times in the dataset, where the species with a short germ line generation time assumes a basal position in the phylogeny (Ohta 1993; Moreira, Philippe 2000). A number of approaches have been explored to systematically deal with fast evolving taxa the most popular include, (1) reconstructing the phylogeny based on slow evolving sites (applied here), (2) increasing the sample size, this is based on the assumption that increasing the sample size actually increases the number of slowly evolving positions, (3) decreasing the distance to the outgroup, and (4) using more accurate models of sequence change such as covarion derivatives.

The gene tree - species tree reconciliation analysis has verified the duplication pattern amongst the MHP. However, it is believed that current methods of reconciliation such as the one used here may be biased towards inferring excess gene duplication and differential loss events, as is the case here. The method only considers the topology and not the corresponding alignment or any rate heterogeneity that may exist (Page, Cotton

2002). It should also be highlighted that the variation of the "*Slow-Fast"* method employed here is an approximate method for a complex evolutionary dynamic and is not without its limitations.

Using this fully resolved phylogeny, positively selected sites have been identified, through the use of Bayesian estimation, unique to all four MHP; MPO, EPO, LPO and TPO. The majority of these sites are in close proximity to catalytically important residues, suggesting that they may potentially be linked to functional shifts across the MHP. The conserved proximal histidines in close proximity to sites under positive selection in MPO, EPO and LPO are crucial in preserving the redox properties of the heme iron for catalysis (Furtmüller et al. 2006). The conserved distal histidines, also shown here to be in the vicinity of positively selected sites, act as both proton acceptors and donor to oxygen during the formation of Compound I, which is an integral step in the peroxidase pathway (Furtmüller et al. 2006). A number of sites identified under positive selection are located in disulphide bond regions, which are believed to be crucial to the structure and function of a protein. Disruption of such regions can be detrimental to the enzymatic stability and activity (Rietsch, Beckwith 1998; Grebski, Peterson, Medici 2001). In particular, six sites pertaining to the LPO family are linked to the same disulphide bond. This strongly suggests that these sites are associated with the unique function of LPO as they are not present in the two closely related families MPO and EPO. In the TPO analysis the majority of the sites with highest probability of being positively selected are located in exon 8 of the protein. Deletion of exon 8 results in misfolding of the TPO protein (Ferrand, Le Fourn, Franc 2003). Exon 8 is also believed to be part of TPOs catalytic centre (exons 8, 9 and 10) (Ambrugger et al. 2001). TPO functional defects are strongly associated with TIOD and several deleterious mutations within this catalytic region have been reported, (Ambrugger et al. 2001; Ferrand, Le Fourn, Franc 2003; Rodrigues et al. 2005). One of the positively selected sites in TPO is associated directly with an inherited deficiency disorder (Rodrigues et al. 2005).

The detailed *in silico* site directed mutagenesis of the positively selected sites for MPO, EPO and LPO has shown that mutating these positions from their positively selected

amino acid state to an alternative ancestral state results in loss/gain of hydrogen bonds between alternative amino acid positions for other sites in particular in the heme binding region of the respective structures. The sites that have been identified as positively selected in the MHP have played a major role in the functioning of these enzymes as evidenced by mutational studies, proximity to active sites and catalytic residues, and inherited disorders.

The results of this study show for the first time from molecular sequence data (i) how this medically important group of enzymes are related to each other, and (ii) suggest that following gene duplication, positive selection has led to the functional diversity observed for the MHP.

In order to determine if these predictions from the evolutionary analysis of the MHP phylogeny were robust, it was necessary to examine these positively selected sites in more detail. To this end, site directed mutagenesis of a small number of sites was carried out, the results of this analysis are detailed in the following results chapter (chapter3).

Chapter 3

In vitro study of positively selected sites in the human MPO enzyme.

The following study was carried out in collaboration with Prof. William Nauseef at the Iowa Inflammation Program, Department of Medicine, University of Iowa and Veterans Affairs Medical Center, Iowa, USA.

3.1 Introduction

Myeloperoxidase (MPO) (EC 1.11.1.7) is a member of the hemoprotein family known as the mammalian heme peroxidases (MHP) (Figure 3.1) (Chapter 2). This homodimeric heme-containing protein is found predominantly in the azurophilic granules of monocytes and neutrophils (polymorphonuclear leukocytes (PMNs)) where its function is critical in the oxygen-dependent innate immune responses of phagocytotic cells (Klebanoff 1970; Klebanoff 1991). All members of the MHP participate in both peroxidation and halogenation cycles, whereby compound I, produced by reaction with H_2O_2, catalyzes one and two electron oxidations, respectively. However, MPO is unique amongst the MHP in its capacity to catalyze the two-electron oxidation of chloride at physiologic pH, thereby generating hypochlorous acid (HOCl). MPO-dependent production of HOCl in the neutrophil phagosome supports potent oxidant killing of bacteria, fungi and tumour cells (Clark, Szot 1981; Stendahl et al. 1984; Koeffler, Ranyard, Pertcheck 1985; Nauseef 1986; Nauseef, Olsson, Arnljots 1988; Dale, Boxer, Liles 2008). In addition, MPO-derived oxidants contribute to the initiation and propagation of inflammatory diseases such as atherosclerosis, degenerative disorders, and autoimmune syndromes, adding to the medical interest in MPO biology (Klebanoff 2005).

The clinical importance of MPO in human immunity is highlighted by inherited MPO deficiency, first observed in the mid-1900s by P. and A. Alius (1954) and Grignaschi *et al*. (Grignaschi et al. 1963). Undritz first assigned the term Alius-Grignaschi Anomaly to the observed condition that is now commonly known as MPO deficiency (Undritz 1966). MPO-deficient neutrophils have varied degrees of impaired microbicidal capacity, with a reduced activity against selected bacteria and an inability to kill several species of *Candida* (Lehrer, Cline 1969; Lehrer, Hanifin, Cline 1969; Klebanoff 1970; Klebanoff 2005). Studies on known causative mutations of MPO deficiency provide invaluable insights into the consequences of genetic abnormalities/mutations on protein function. Several different genotypes of inherited MPO deficiency have been identified, including

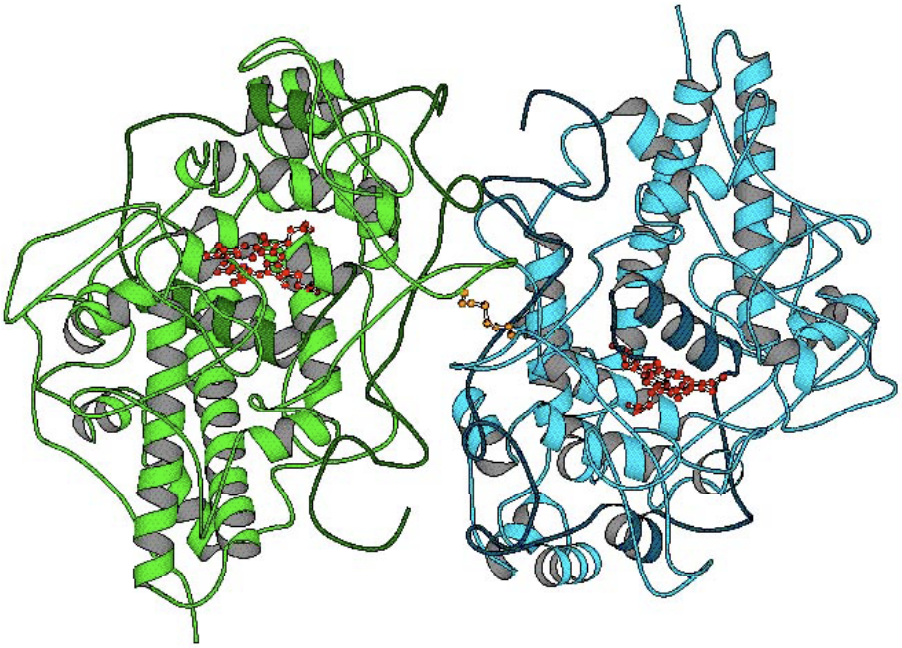


Figure 3.1. Human myeloperoxidase. The two identical monomers of the homodimer are seen in green and blue, the incorporated heme group in red and the disulphide bridge linking the two monomers in orange. Adapted from http://metallo.scripps.edu/PROMISE/.

six genetic mutations that have been demonstrated to exhibit aberrant biosynthesis of MPO: Y173C, M251T, R499C, G501S, R569W and a 14-base deletion in exon 9 (Nauseef, Brigham, Cogley 1994; Nauseef, Cogley, McCormick 1996; Romano et al. 1997; DeLeo et al. 1998; Ohashi et al. 2004; Persad et al. 2006).

Six mutations have also been identifed in an Italian study of MPO deficient patients that are believed to be associated with the deficiency by affecting the structural integrity of the enzyme. These mutations are: A332N, D371G, L572W, W643R, an adenine deletion within exon 3 and a mutation within the 3' splice site of intron 11 (Marchetti et al. 2004b; Marchetti et al. 2004a). All of the above mutations associated with MPO deficiency are summarised in Table 3.1. The missense mutations Y173C, R499C, R569W, and G501S result in mutant precursors that fail to undergo proteolytic maturation and consequently lack significant peroxidase activity (Nauseef, Cogley, McCormick 1996; DeLeo et al. 1998; Goedken et al. 2007). Further to this, the missense mutation, M251T, results in inefficient processing into mature MPO with a significant reduction in enzymatic activity (Romano et al. 1997).

Human MPO is encoded by a single gene located on the long arm of chromosome 17. Its biosynthesis takes a series of modifications involving the incorporation of heme which leads ultimately to the production of an active homodimer (see Figure 3.2 for a schematic of normal human MPO biosynthesis). The primary 80 kDa translation product, preproMPO, undergoes glycosylation producing the enzymatically inactive apoproMPO. Subsequent insertion of heme into the peptide backbone of apoproMPO generates the active 90 kDa heme-containing precursor, proMPO. Heme incorporation during MPO biosynthesis is a prerequisite for its activity. Two molecular chaperones, calreticulin (CRT) and calnexin (CLN), are associated with the biosynthesis of MPO precursors in the endoplasmic reticulum (ER). These ER molecular chaperones play a role in correct folding of glycoproteins (Helenius *et al*. 1997). Both chaperones interact with apoproMPO, perhaps facilitating correct folding of the precursor. It has been suggested that CRT participates in the incorporation of the heme group into apoproMPO to generate the active proMPO precursor in the ER, with subsequent exit into the Golgi

Table 3.1: Mutations associated with MPO deficiency.

Known Causative Mutation	**Effect**	**Reference**
R569W	Maturational arrest	Nauseef et al. 1994; Nauseef et al. 1996
M251T	Low enzyme activity	Romano et al. 1997
14-base deletion exon 9	Aberrant mRNA splicing	Romano et al. 1997
Y173C	Malfolded protein	DeLeo et al. 1998
G501S	Aberrant processing	Ohashi et al. 2004; Goedken et al. 2007
R499C	Aberrant processing	Persad et al. 2006; Goedken et al. 2007
Putative MPO Deficiency Mutations	**Effect**	**Reference**
A332V	Potential structural changes	Marchetti et al. 2004a; Marchetti et al. 2004b
D371G	Potential structural changes	"
L572W	Potential structural changes	"
W643R	Potential structural changes	"
Adenine deletion exon 3	Frame shift resulting in a premature stop codon in pro-peptide	"
Mutation in 3' splice site intron 11	Potential aberrant mRNA splicing	"

All mutations have been identified by genotyping. The 'Known Causative Mutations' have been characterised *in vitro,* potential effects of "Putative MPO Deficiency Mutations' have been suggested.

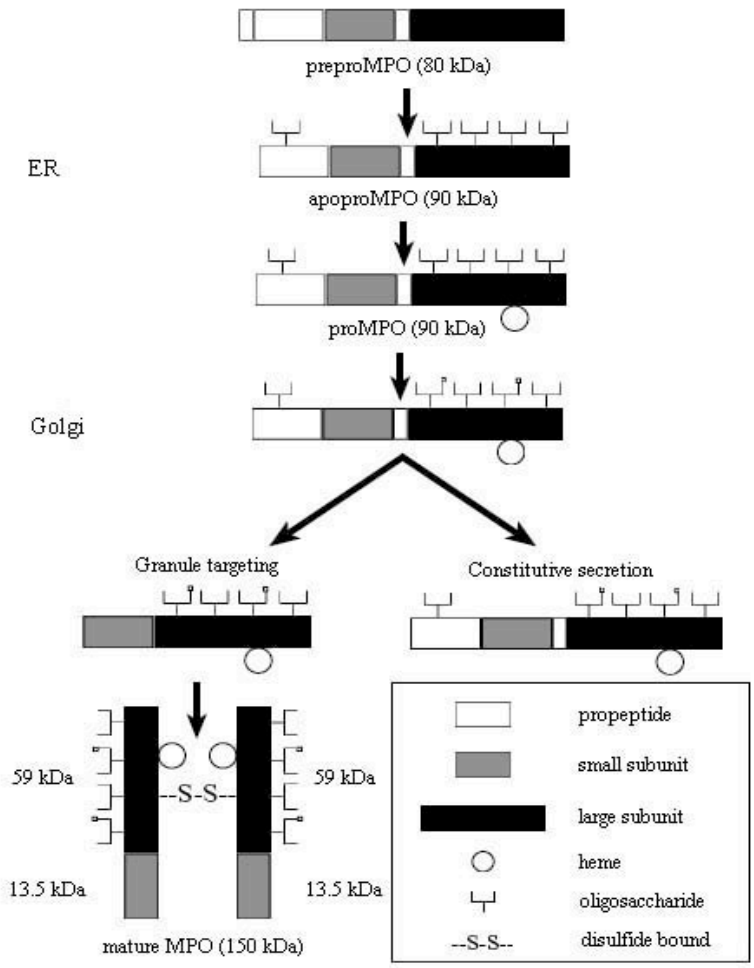


Figure 3.2. Normal MPO biosynthesis. A schematic of the normal processing and maturation of human MPO. The synthesised 80 kDa primary translation product (preproMPO) consists of a propeptide region and a small and large subunit. Glycosylation of preproMPO yields the inactive 90 kDa precursor (apoproMPO). ApoproMPO associates with molecular chaperons, calreticulin and calnexin, in the endoplasmic reticulum (ER), resulting in the incorporation of the essential heme co-factor, generating the active precursor (proMPO). Upon heme acquisition, proMPO exits to the Golgi for further processing and granule targeting. A short-lived 75 kDa intermediate (lacking the propeptide region) is subsequently cleaved into a two-subunit (13.5 and 59 kDa) monomer form. A 150 kDa homodimer (mature MPO) is formed with each monomer linked by a disulphide bond. Adapted from (Olsson, Bulow, Hansson 2004).

(Nauseef, McCormick, Clark 1995; Hansson, Olsson, Nauseef 2006). proMPO undergoes further processing and dimerises to produce the mature homodimer. The two identical monomers are linked by a disulphide bridge at position C319 resulting in mature dimeric MPO of approximately 150 kDa (Hansson, Olsson, Nauseef 2006). Each monomer consists of a heavy and light subunit of 59 kDa and 13.5 kDa respectively.

In this thesis an evolutionarily informed approach to identify and prioritise functionally essential sites for targeted mutagenesis has been used. It is generally accepted that the primary activity of MPO is the generation of HOCl, a capacity unique to MPO among its MHP counterparts. Previously, it has been shown that positive selective pressure on the four main members of the MHP family [MPO, eosinophil peroxidase (EPO), thyroid peroxidase (TPO) and lactoperoxidase (LPO)] contributed to the observed functional diversity in these enzymes (Chapter 2). Amino acid residues have been identified that are unique to MPO and are predicted to have played an important role in the evolution of its specific function. The correlation between positive selective pressure, as measured by nucleotide substitutions and codon based models of evolution (Yang 1997; Yang et al. 2000), and functional divergence has rarely been investigated at both the genotypic and phenotypic level.

To test the hypothesis that the identified residues reflect positively selected amino acid residues essential for the unique function of MPO, a directed mutagenesis approach was applied followed by biochemical analyses of the effects of mutating these residues (Nauseef, Cogley, McCormick 1996; DeLeo et al. 1998; Goedken et al. 2007). Site directed mutagenesis was performed on four sites, R80, N496, Y500 and L504, identified from the *in silico* evolutionary study in chapter 2. These sites were chosen based on: (i) the confidence score from the *in silico* predictions, (ii) their spatial relationship with the proximal heme ligand, His 502, and (iii) their proximity to R499C and G501S, mutations known to cause MPO deficiency. Position R80 is located in the propeptide region and not in the heme pocket. Based on the resolved phylogenetic history of the MHP, R80, N496, Y500 and L504 sites were mutated *in vitro* to their more ancestral state, methionine (M), phenylalanine (F), F and threonine (T),

respectively (R80M, N496F, Y500F and L504T). Site directed mutagenesis at positons 80 and 496 was unsuccessful; therefore, only mutations at positions 500 and 504 were further analysed. This study has demonstrated that the substitutions Y500F and L504T, independently and in combination (*i.e.* double mutant), disrupted normal MPO biosynthesis and severely decreased enzymatic activity. These findings indicate that these residues are indeed closely associated with MPO-specific enzymatic function and demonstrate the tangible link between *in silico* evolutionary predictions and protein function.

3.2 Methodology

3.2.1 Biological Materials

The following tables detail the biological materials used in this study:

Table 3.2: Mammalian Cell lines used in this study.

Cell Line	Description
HEK 293*	Human embryonal kidney cells 293 (HEK) lacking endogenous MPO
HEK MPO	HEK cell line stably transfected with pMPO, capable of expressing recombinant hMPO
HEK Y500F	HEK cell line stably transfected with pMPO-Y500F, capable of expressing variant recombinant hMPO
HEK L504T	HEK cell line stably transfected with pMPO-L504T, capable of expressing variant recombinant hMPO
HEK Y500F/L504T	HEK cell line stably transfected with pMPO-Y500F/L504T, capable of expressing variant recombinant hMPO

*Obtained from American Type Culture Collection (Manassa, VA), ATCC CRL-1573, All other cell lines were generated in this study.

Table 3.3: Antibodies used in this study.

Antibody	Name	Description	Source
Rabbit anti-MPO (primary)	Rabbit anti-MPO	Rabbit polyclonal anti-sera directed against hMPO	Professor William M. Nauseef, Iowa Inflammation Program, University of Iowa, USA
Anti-rabbit HRP (secondary)	Anti-rabbit IgG (whole molecule) – peroxidase antibody produced in goat	A goat anti-rabbit antibody that reacts with rabbit IgG, and is conjugated to peroxidase	Sigma-Aldrich
Mouse anti-β-actin (primary)	Monoclonal anti-β-actin produced in mouse, clone AC-15	An IgG purified monoclonal that recognises an N-terminal peptide of actin protein	Sigma-Aldrich
Anti-mouse AP (secondary)	Goat anti-mouse IgG, Alkaline Phosphatase (AP) conjugate.	A goat anti-mouse antibody that reacts with mouse IgG, and is conjugated to AP	Promega

Table 3.4: Bacterial strain used in this study.

Bacterial Strain	Genotype	Source
E. coli XL10-Gold	Tet^r D(mcrA)183 D(mcrCB-hsdSMR-mrr)173 endA1 supE44 thi-1 recA1 gyrA96 relA1 lac Hte F'[proAB lacI ^{q}ZDM15 Tn10 (Tet^r) Amy Cam^r]	Stratagene

Table 3.5: Plasmids used in this study.

Plasmid	Description	Source
pcDNA3.1(-) Neo	Mammalian high-level constitutive expression vector with CMV enhancer-promoter. Ampicillin and geneticin™ resistance for selection in *E. coli* and mammalian cells respectively.	Invitrogen
pcDNA-MPO	pcDNA3.1(-) Neo with hMPO gene present.	Prof. Nauseef, Iowa, USA
pMPO	pcDNA3.1(-) Neo with hMPO gene present, no missense mutations present.	Prof. Nauseef, Iowa, USA
pMPO-Y500F	pMPO with tyrosine at position 500 replaced with phenylalanine	This study
pMPO-L504T	pMPO with leucine at position 504 replaced with threonine	This study
pMPO-Y500F/L540T	pMPO with tyrosine at position 500 replaced with phenylalanine and leucine at position 504 replaced with threonine	This study
pUC18	Transformation control vector	Stratagene
pWhitescript™	Mutagenesis control vector	Stratagene

Table 3.6: Oligonucleotides used in this study.

Target	Primer Sequence
pcDNA3.1 For*	5'-GGCTAACTAGAGAACCCACTG-3'
pcDNA3.1 Rev*	5'-GGCAACTAGAAGGCACAGTC-3'
Mid For*	5'-CCAGCTGTTGGACCACGACCTCG-3'
Mid Rev*	5'-GACTCAGTGGACCCACGCATCGCC-3'
R80M For†	5'-CCTACAAGGAG**AT**GCGGGAAAGC-3'
R80M Rev†	5'-GCTTTCCCGC**AT**CTCCTTGTAGG-3'
N496F For†	5'-CGTCTTCACC**TT**TGCCTTCCGC-3'
N496F Rev†	5'-GCGGAAGGCA**AA**GGTGAAGACG-3'
Y500F For†	5'-CCTTCCGCT**TT**GGCCACACCC-3'
Y500F Rev†	5'-GGTGTGGCC**AA**AGCGGAAGGC-3'
L504T For†	5'-GGCCACACC**AC**CATCCAACCC-3'
L504T Rev†	5'-GGGTTGGATG**GT**GGTGTGGCC-3'

*Sequencing primers †Mutagenic primers

Letters in boldface indicate the specific nucleotide base changes required to alter the amino acid coded for.

3.2.2 DNA Manipulation

3.2.2.1 Plasmid Preparation

A single *E. coli* XL-10 Gold plasmid-bearing colony was picked from a LB agar plate containing 100 µg/mL ampicillin and used to inoculate 10 mL of LB ampicillin (100 µg/mL) liquid broth. This was incubated overnight (>16 h) at 37 ºC and 220 rpm. The plasmid of interest was purified using the Sigma-Aldrich GenElute™ Plasmid Miniprep Kit. Overnight culture (1.5 mL) was harvested by centrifugation at 16,000 x *g* for 1.5 min. Supernatant was discarded and the cell pellet was resuspended in 200 µL Resuspension Solution by pipetting. Lysis Solution (200 µL) was then added and the contents were mixed by inversion and allowed to clear for up to 5 min. Neutralisation Solution (350 µL) was added and the contents were mixed by inversion. Cell debris was pelleted for 10 min at 16,000 x *g*. The binding column was prepared by passing 500 µL of Column Preparation Solution through at 16,000 x *g* for 1 min and discarding the flow through. Plasmid DNA was bound to the column by passing the cleared lysate through at 16,000 x *g* for 1 min, discarding the flow through. Contaminants were removed by washing the DNA-bound column with 500 µL of Optional Wash Solution at 16,000 x *g* for 1 min. Wash Solution (750 µL) was then added and centrifuged for 1 min 16,000 x *g*. The flow through was discarded and the column was centrifuged for 1 min at 16,000 x *g* to dry it. The dry DNA-bound column was transferred to a new collection tube and the purified plasmid DNA was eluted from the column by passing 50 µL of Elution Solution through at 16,000 x *g* for 1 min.

3.2.2.2 Restriction digestion of DNA

Restriction analysis was carried out to cut specific fragments from a plasmid. Restriction digest patterns were predicted for DNA sequences using NEBCutter online tool (http://tools.neb.com/NEBcutter2). All restriction enzymes used were supplied with 10X concentration of incubation buffers (working concentration 1X). DNA digestions were

performed according to manufacturer's instruction (New England Biolabs) and incubated for up to 3 h at the optimum restriction enzyme temperature.

3.2.2.3 Ligation of DNA

Equimolar amounts of vector and insert DNA (1 µg) were ligated overnight at 17 ºC or for 3 h at 22 ºC in Invitrogen's T4 DNA ligase buffer (10 units of T4 DNA ligase/mL) in a total volume of 10 µL. Ligated samples were then incubated at 70 ºC for 10 min to inactivate the ligase and were then transformed immediately or stored at -20 ºC until required.

3.2.2.4 Site directed mutagenesis

pMPO was used as template DNA for mutagenesis. The mutagenesis PCR reaction mix (Table 3.7) was designed according to the Stratagene QuickChange® II XL Site-Directed Mutagenesis Kit guidelines and was utilised for all site-specific mutagenesis.

Table 3.7: Mutagenesis PCR reaction mix.

Component	Volume
sH_2O	34 µl
10X reaction buffer	5 µl
pMPO (10 ng/µL)	2 µl
Forward primer (100 pmole)	2 µl
Reverse primer (100 pmole)	2 µl
dNTP mix (10 mM)	1 µl
Quick solution	3 µl
Pfu turbo polymerase (2.5 U/µL)	1 µl

All mutagenesis PCR reactions were performed in a Thermo Electron Corporation PX2 Thermal Cycler under the cycle conditions in Table 3.8.

Table 3.8: Mutagenesis PCR programme.

Segment	Cycles	Temperature	Time
1 (Denaturation)	1	95 ºC	1 min
2 (Cycling: Denaturation-Annealing-Elongation)	18	95 ºC	50 sec
		65 ºC	50 sec
		68 ºC	7 min 40 sec
3 (Elongation)	1	68 ºC	7 min

The amplification products were digested with 1 µL of *Dpn* I (10 U/µL) for 1 h at 37 ºC to digest parental DNA. *Dpn* I-treated DNA was transformed into Stratagene XL10-Gold ultracompetent cells.

3.2.2.5 Transformation

Stratagene XL10-Gold ultracompetent cells were transformed by heat shock. All media components, complex buffers and solutions are described in the Appendix. A microfuge tube of ultracompetent cells was allowed to thaw on ice. Thawed cell susspension (45 µL) was aliquoted into a 14 mL BD Falcon tube. β-Mercaptoethanol (2 µL) was added and mixed by swirling every 2 min for 10 min on ice. *Dpn* I-DNA (2 µL) was added, mixed by swirling and incubated on ice for 30 min. The sample was then incubated at 42 ºC for 30 sec and immediately incubated on ice for a further 2 min. Preheated (42 ºC) NZY+ medium (500 µL) was added, the contents mixed by swirling and incubated at 37 ºC for 1 h at 220 rpm. Experimental sample (250 µL) was then plated on LB agar containing 100 µg/mL ampicilin. Mutagenic and transformation control samples (250 µL each) were then plated on LB agar containing 100 µg/mL ampicillin, 80 µg/mL X-gal and 20 mM IPTG. All plates were incubated overnight at 37 ºC. Resultant experimental colonies were used to prepare broth cultures for plasmid DNA preparations.

3.2.2.6 Agarose gel electrophoresis

Electrophoresis through agarose gel is commonly used to separate, identify and sometimes purify DNA fragments. Agarose gel was prepared by boiling 0.7-1 % w/v agarose in 1 X TAE (Appendix) until the solution became translucent. The solution was allowed to cool and then poured into the mould apparatus; combs were inserted and the solution was allowed to solidify. The gel was placed in the gel box apparatus, the comb was removed and 1X TAE buffer was added to fill the electrode chamber and gel box 1 cm above the level of the gel. DNA samples containing loading buffer (Appendix) were loaded into the wells and the gels were electrophoresed at 100 Volts for approx. 1 h. DNA ladders of 1 kb were also used. Gels were stained by immersing in a bath of ethidium bromide (Appendix) for 25 min and destained with water for 10 min. The gels were visualised with a UV transilluminator and photographed using a UV image analyser.

3.2.2.7 DNA quantification and sequencing

Sequencing of plasmid DNA was performed to verify (i) the presence of desired mutations, (ii) the absence of unintentional mutation, and (iii) "in-frame" insertion of DNA fragments following cloning. Plasmid DNA was prepared as per Section 3.2.2.1. A sample of plasmid DNA (1 µL) was applied to a NanoDrop Spectrophotometer ND-100 and the concentration of DNA in the sample was recorded. A sample of plasmid (15 µL; 50-100 ng/µL) was sent to Eurofins/MWG-Biotech, London, United Kingdom or Integrated DNA Technologies, Iowa, U.S.A for sequencing using the appropriate sequencing primers in Table 3.6.

3.2.3 Cell Culture Methods

All tissue culture techniques were performed in a sterile environment using a Holten laminar flow cabinet (model 1.2). All medium components and supplements are given in the Appendix.

3.2.3.1 Culture of adherent cells

All HEK cell lines were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10 % (v/v) fetal bovine serum, 100 U/mL penicillin, 100 µg/mL streptomycin, 100 mM HEPES and 2 mM L-glutamine and selected using 600 µg/mL G-418 sulphate (with the exception of the wild type HEK 293 cell line). HEK cells were seeded into 25 cm^2 and 75 cm^2 tissue culture flasks. All cell lines were incubated in a humidified 5 % CO_2 atmosphere at 37 ºC in a cell culture incubator. Adherent cells were detached by trypsinisation or basic dislodgement by tapping flask sides. The cell suspension was decanted into a sterile centrifuge tube and centrifuged at 200 x *g* for 6 min. Cells were resuspended in supplemented medium at 2 to 5 x 10^5 cells/mL, using 5 mL per 25 cm^2 and 15 mL per 75 cm^2 flasks, and incubated as above.

3.2.3.2 Cell counts

Viable cells were counted using ethidium bromide/acridine orange (EB/AO) staining, where viable cells fluoresce green and apoptotic cells have red-orange fluorescence. Viable cells were counted on a haemocytometer slide. Cell sample (20 µL) was added to 80 µl of 1X EB/AO solution and mixed by pipetting. Ten microlitres of this mixture were added to the counting chamber of the haemocytometer and cells were visualized using a Nikon Eclipse e200 fluorescent microscope. Cell counts were expressed as the number of cells per mL.

3.2.3.3 Transient transfection

Mutagenic plasmid (12 µg) was diluted in cell growth medium containing no serum, protein or antibiotics to a total volume of 150 µL. Qiagen® PolyFect Transfection Reagent (115 µL) was added and the sample was mixed by pipetting and incubated at room temperature for 10 min. Spent growth medium from 40-80 % confluent HEK 293 cells (75 cm^2 flask) was gently aspirated. Cells were gently washed once with 15 mL of 1X PBS without dislodging from flask surface. Fresh growth medium (7 mL) containing supplements was added. Fresh supplemented medium (1 mL) was added to the

transfection sample. This mixture was then added to the HEK 293 cells and incubated overnight in a humidified 5 % CO_2 atmosphere at 37 ºC in a cell culture incubator.

3.2.3.4 Stable transfection

Stable cell lines were established by transfection (Section 3.2.3.3) followed by selection with G-418 sulphate 48 h later. Within 3-6 weeks, G-418 sulphate (400 up to 600 μg/mL) resistant cells grew out.

3.2.4 Protein Analysis

3.2.4.1 3D modeling and *in silico* mutational analysis

Homology modeling and *in silico* mutational analysis of two previously identified myeloperoxidase specific positively selected sites (positions Tyr500 and Leu504) were performed and the impact, *in silico*, of two single mutations (Y500F and L504T) and the double mutant (Y500F/L504T) on hydrogen bonding was assessed as described in Chapter 2.

3.2.4.2 Pulse-chase analysis of MPO biosynthesis

Radioactive pulse-chase experiments can be utilised to track the progression of cellular processes over time. All medium components, complex buffers and solutions are described in the Appendix. Spent medium from overnight transient/stable transfectants was gently aspirated. Fresh cell growth medium (10 mL) containing serum, antibiotics and 2 μg/mL hemin was added and incubated overnight at 37 ºC and 5 % CO_2. Spent medium was gently aspirated. Cells were washed twice with 10 mL of PBS. Methionine-free cell growth medium (5 mL) containing 2 μg/ml hemin was added and incubated for 1 h at 37 ºC and 5 % CO_2. Cells were pulse-labelled with 25 μCi/mL ^{35}S (specific activity: 1,500 Ci/mmol) for 1 h at 37 ºC and 5 % CO_2. Cells were then chased for 20 h at 37 ºC and 5 % CO_2 by the addition of 50 μL of 100 mM cold methionine. Cells and

medium were collected by centrifugation for 6 minutes at 200 x *g*. Inhibition buffer (26 μL) was added to 1 mL of medium and stored on ice for 20 minutes. Cell pellets were resuspended in 400 μL Leu-pep buffer, 4 μl 100 mM PMSF-isopropanol was added and the mixture was stored on ice for 20 min. Non-immune serum (7 μL) was added to 200 μL of cell mix and 1 mL of supernatant mix. Samples were tumbled for 30 min at 4 ºC. Pansorbin (50 μL) was added and samples were tumbled for 30 min at 4 ºC. Pansorbin pellets were collected by centrifugation for 4 min at 16,000 x *g* and discarded. Dilution buffer (300 μL) was added to the cell supernatant. Anti-MPO (7 μL) was added to each medium supernatant and buffered cell supernatant and tumbled for 4 h at 4 ºC. Pansorbin (50 μL) was added and samples were tumbled overnight at 4 ºC. Samples were collected at 16,000 x *g* for 30 sec; supernatant was discarded. Pellets were washed with 900 μL 0.5 % (v/v) Triton 100/TBS and collected at 16,000 x *g* for 30 sec; supernatant was discarded. Pellets were washed with 900 μL 2 M Urea/0.5 % (v/v) Triton 100/TBS and collected at 16,000 x *g* for 30 sec, supernatant was discarded. Pellets were washed with 900 μL 1 mg/ml BSA/0.5 % (v/v) Triton 100/TBS and collected at 16,000 x *g* for 30 sec; supernatant was discarded. Pellets were washed with 900 μL TBS and collected at 16,000 x *g* for 30 sec; supernatant was discarded. Pellets were resuspended in 65 μL SDS sample buffer. Samples were heated to 100 ºC for 4 min, centrifuged at 16,000 x *g* for 4 min and the supernatant was analysed by SDS-PAGE. Gels were washed in destaining solution for 20 min, followed by washing in water for 20 min and finally by staining in 1 M sodium salicylate, pH 6.2. Stained gels were then dried under vacuum at 80 ºC. This was followed by autoradiography by exposing dry gels to X-ray film for up to 4 days and developing. MPO-related protein was quantified by densitometry using a PhosphorImager (Typhoon 9410, Amersham Biosciences).

3.2.4.3 SDS-polyacrylamide gel electrophoresis

Polyacrylamide gel electrophoresis (PAGE) is used to separate proteins based on their electrophoretic mobility. PAGE is generally performed in the presence of sodium dodecylsulphate (SDS), a negatively charged detergent that binds to all types of protein

molecules. SDS-protein complexes migrate through the polyacrylamide gels based on the size of the polypeptide.

SDS-PAGE was performed using a 10 % (v/v) resolving gel and a 5 % (v/v) stacking gel. Gel components, buffers and solutions were prepared as detailed in the Appendix. An ATTO protein gel electrophoresis apparatus was used in this study. Glass plates, gasket and comb were washed with detergent, rinsed with dH_2O and wiped in one direction with 100 % (v/v) ethanol. A gasket was placed around the ridged plate and the plates were assembled and secured with clamps. The resolving gel was poured to below 2.5 cm from the top of the plates and overlaid with 100 % (v/v) ethanol until the gel polymerised. The ethanol was removed and the stacking gel was poured to the top of the plates. A comb was inserted and the gel was allowed to polymerise. The electrophoresis tank was filled to approx. 5 cm from the base of the tank with 1X running buffer. The clamps and gasket were removed from the polymerised gel and the gel was lowered into the buffer, excluding any air bubbles from the base of the gel. The gel was secured in place using pressure plates. The tank was filled with 1X running buffer and the comb was removed. Sample wells were rinsed with 1X running buffer to remove any un-polymerised gel. Samples were loaded and electrodes attached. The gel was electrophoresed at a constant current of 30 mA per gel until the blue dye front reached the bottom of the gel. The glass plates were removed and the gel was subjected to staining (Section 3.2.4.2 or Section 4.2.4.7 (Chapter 4 analyses)) or western blotting (Section 3.2.4.4).

3.2.4.4 Western blotting

3.2.4.4.1 Preparation of MPO protein for western blotting

All medium components, complex buffers and solutions are described in the Appendix. Supplemented cell growth medium containing 2 μg/mL hemin was added to 60-70 % confluent cells and incubated for 48 h at 37 ºC and 5 % CO_2. Cells were collected by centrifugation at 200 x *g* for 6 min, spent medium was removed and 1000 μL of packed cells were stored at 4 ºC. Cell pellets were resuspended in 10-15 mL 1X PBS and

centrifuged at 200 x *g* for 6 min; the supernatant was discarded. This step was repeated and pellets were resuspended in the same volume of 1X PBS. Cells were counted as per Section 3.2.3.2. At least 5 million cells/mL were collected by centrifugation at 200 x *g* for 6 min. The cell pellet was resuspended in 1000 µL 1X PBS and transferred to a microcentrifuge tube. Cells were pelleted by centrifugation at 200 x *g* for 6 min. Packed cell pellets were resuspended by vortexing in 20 volumes of ProteoJET™ Mammalian Cell Lysis Reagent (containing protease inhibitors) to 1 volume of packed cells. The mixture was tumbled at room temperature for 10 min. The cell lysate was clarified by centrifugation at 16,000 x *g* for 15 min and was either used immediately or stored at -80 ºC. An equal volume of 2X SDS loading buffer was added to the clarified lysate and supernatant samples. The mixture was boiled for 5 min and the cell debris was pelleted by centrifugation at 16,000 x *g* for 10 min and discarded. The samples were then subjected to SDS-PAGE as per section 3.2.4.3 or stored at -20 ºC.

3.2.4.4.2 Immunological probing

Electrophoretically separated proteins were transferred from the polyacrylamide gel to a nitrocellulose membrane using the Invitrogen iBlot™ Dry Blotting System. MPO-related proteins were detected on the blot membrane using the Millipore SNAP i.d.™ Protein Detection System and subsequent chemiluminescence. The transferred membrane was probed with a primary antibody specific to MPO or β-actin. The bound primary antibody was then probed with a secondary antibody and detected by chemiluminescence or colorimetric detection.

The iBlot™ Gel Transfer Device was set up as per Figure 3.3 below. Any air bubbles were removed from the bottom stack (Anode + buffer gel + blotting membrane) using the de-bubbling roller. The pre-run gel was overlaid on top of the bottom stack and air bubbles were removed. Filter paper was soaked in dH_2O and placed on top of the gel, expelling any bubbles present. The top cathode stack (buffer gel + cathode) was then placed over the bottom layers. Protein transfer to the membrane was achieved by applying 15 V for 7 min using the iBlot™ Dry Blotting System.

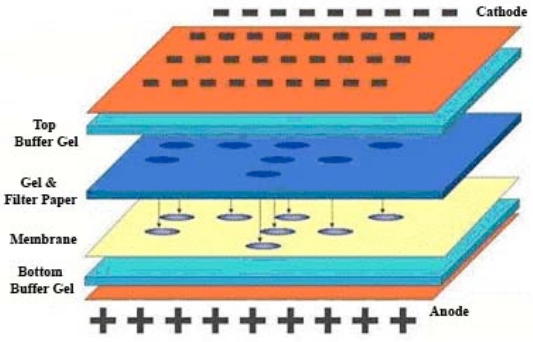


Figure 3.3. iBlot™ Gel Transfer Device. Adapted from http://www.invitrogen.com.

The transferred membrane was then stained with Ponceau S stain to ensure that uniform transfer of proteins to the membrane had been achieved. Briefly, the transferred membrane was immersed in 10 mL Ponceau S stain and incubated at room temperature for 5 min with constant agitation. Proteins were visualised as red bands. The membrane was then washed several times with dH_2O until the stain had been washed away.

The transferred membrane was then blocked and probed with specific antibodies using the Millipore SNAP i.d.™ System. The inner surface of the blot holder was dampened with dH_2O. The transferred membrane was soaked in dH_2O and placed protein side down in the blot holder. Any air bubbles were removed using the de-bubbling roller. The spacer layer was then placed on top of the membrane and rolled to ensure contact between spacer and membrane. The blot hold was then closed and placed in the SNAP i.d. system. The membrane was blocked by passing 15 mL of blocking buffer (Appendix) over the membrane under vacuum. The membrane was then removed from the blot holder and incubated overnight at 4 ºC in 10 mL of the appropriate primary antibody (Table 3.3) with constant agitation. The membrane was inserted into the blot holder and system as above and the 10 mL of overnight primary antibody was passed through the membrane under vacuum. The membrane was then washed by passing 15 mL TBS-T (Appendix) through under vacuum three times. The washed membrane was then incubated in 10 mL of the appropriate secondary antibody (Table 3.3) for 30 min at room temperature followed by washing as above. For MPO detection the probed membrane was then incubated in 5 mL of SuperSignal® west pico chemiluminescent

substrate (Thermo Fisher Scientific) for 5 min in the dark. The membrane was then covered in plastic wrap. Luminescence was detected, and MPO-related protein was quantified by densitometry, using Syngene's GeneGnome HR Bio-imaging system and GeneTools respectively. For β-actin detection, the probed membrane was then incubated in 5 mL of 5-Bromo-4-chloro-3-indolyl phosphate/Nitro Blue Tetrazolium (BCIP/NBT, Sigma-Aldrich) for 5 min in the dark. This colormetric detection for alkaline phoshatase results in the production of a visible coloured product. β-Actin was used simply as a loading control, so visualization of the coloured product was sufficient and no densitometric analysis was required.

3.2.4.5 Peroxidase activity assay

All medium components, complex buffers and solutions are described in the Appendix. Supplemented cell growth medium containing 2 μg/mL hemin was added to 60-70 % confluent cells and incubated for 48 h at 37 ºC and 5 % CO_2. Cells were collected by centrifugation at 200 x *g* for 6 min and spent medium was removed. Cell pellets were resuspended in 10-15 mL 1X PBS and centrifuged at 200 x g for 6 min; supernatant was discarded. This step was repeated and pellets were resuspended in the same volume of 1X PBS. Cells were counted as per Section 3.2.3.2. At least 5 million cells/mL were collected by centrifugation at 200 x *g* for 6 min. Cell pellet was resuspended in 0.01 % (v/v) Triton X-100/1X PBS to a density of 1 million cells/18 μL and stored on ice. Peroxidase assay was performed in a 37 ºC water bath. Cell sample (18 μL) was added to 3.5 mL of TMB buffer mix. H_2O_2 (0.49 M; 2.1 μL) was added to initiate the reaction; the contents were mixed by vortexing. After 3 min, 100 μL of 0.35 mg/mL catalase was added and mixed by vortexing to stop the reaction and the mixture was placed on ice. Ice-cold 0.2 M acetic acid (3.4 mL) was immediately added. Absorbance was then recorded at OD_{655} on a a Shimadzu UVmini-1240 spectrophotometer.

3.2.4.6 Chlorination activity assay

All medium components, complex buffers and solutions are described in the Appendix. Supplemented cell growth medium containing 2 μg/mL hemin was added to 60-70 % confluent cells and incubated for 48 h at 37 ºC and 5 % CO_2. Cells were collected by centrifugation at 200 x *g* for 6 min and spent medium was removed. Cell pellets were resuspended in 10-15 mL 1X PBS and centrifuged at 200 x *g* for 6 min; supernatant was discarded. This step was repeated and pellets were resuspended in the same volume of 1X PBS. Cells were counted as per Section 3.2.3.2. At least 5 million cells/mL were collected by centrifugation at 200 x *g* for 6 min. The cell pellet was resuspended in 1X PBS to a density of 1 million cells/50 μL. The sample was sonicated on ice at 40 % amplitude for 30 sec with 6 sec pulses using a Branson Digital Sonifer®. Sonicated samples were centrifuged at 16,000 x *g* for 10 min at 4 ºC. The supernatant was stored on ice. The chlorination assay was performed using Invitrogens EnzChek® Myeloperoxidase (MPO) Activity Assay Kit. All kit components were brought to room temperature prior to commencing the assay. Kit MPO standards (0-200 ng/mL) were prepared in 1X PBS. Experimental and standard samples (50 μL each) were added to a 96-well microplate. Fifty microlitres of 2X 3'-(*p*-aminophenyl) fluorescin (APF) working solution was added to all experimental and standard sample wells. The microplate was then incubated in the dark at 37 ºC for 5 min. Fluorescent intensity of each sample was recorded at 485 nm excitation and 530 nm emission on a Perkin Elmer luminescence spectrometer (model LS 50 B).

3.3 Results

3.3.1 Impact of mutation of positively selected sites on the strucural integrity of MPO *in silico*

Performing detailed *in silico* site-directed mutagenesis, Chapter 2 assessed the impact of positively selected sites on the hydrogen bonding around the heme group in MPO and, hence, on the structural integrity of the enzyme. This chapter focuses on just two positions predicted from the previous analyses. These are Tyr 500 and Leu 504. These residues are located up- and down-stream of the proximal heme ligand, His 502. The target mutant amino acid for these two residues was inferred based on the evolutionary relationship of the MHP and not on amino acid properties. Based on this evolutionary analysis, Y500 and L504 were mutated to their respective ancestral residues, phenylalanine (F) and threonine (T), respectively. The effect of the MPO-specific positively selected sites on hydrogen bonding (and, hence the enzyme's structural integrity) was investigated by performing detailed *in silico* site directed mutagenesis. Any resulting changes in a loss/gain of hydrogen bonds, particularly in the heme binding pocket of MPO, were noted.

Data from the 3D structure of the enzyme indicate that Y500 lies between the two main residues associated with MPO deficiency, R499C and G501S (Figure 3.4a). The proximal heme ligand, H502, is connected to R499 via a hydrogen bond, and is directly bound to G501 and T503. Y500 shares putative hydrogen bonds with Y462, A497 and T503. By mutating position 500 to a phenylalanine, it is predicted that the hydrogen bond with Y462 would be lost (see Figure 3.4a). Residue L504 is directly bound to T503 and shares hydrogen bonds with G501 and K556. Mutating leucine at position 504 to threonine (L504T) would likely create an additional hydrogen bond with G501 (see Figure 3.4b). Mutating both Y500 and L504 in combination should not change the outcome of the above predictions; Y500F would result in the loss of a hydrogen bond and L504T in the formation of an additional hydrogen bond (Figure 3.5).

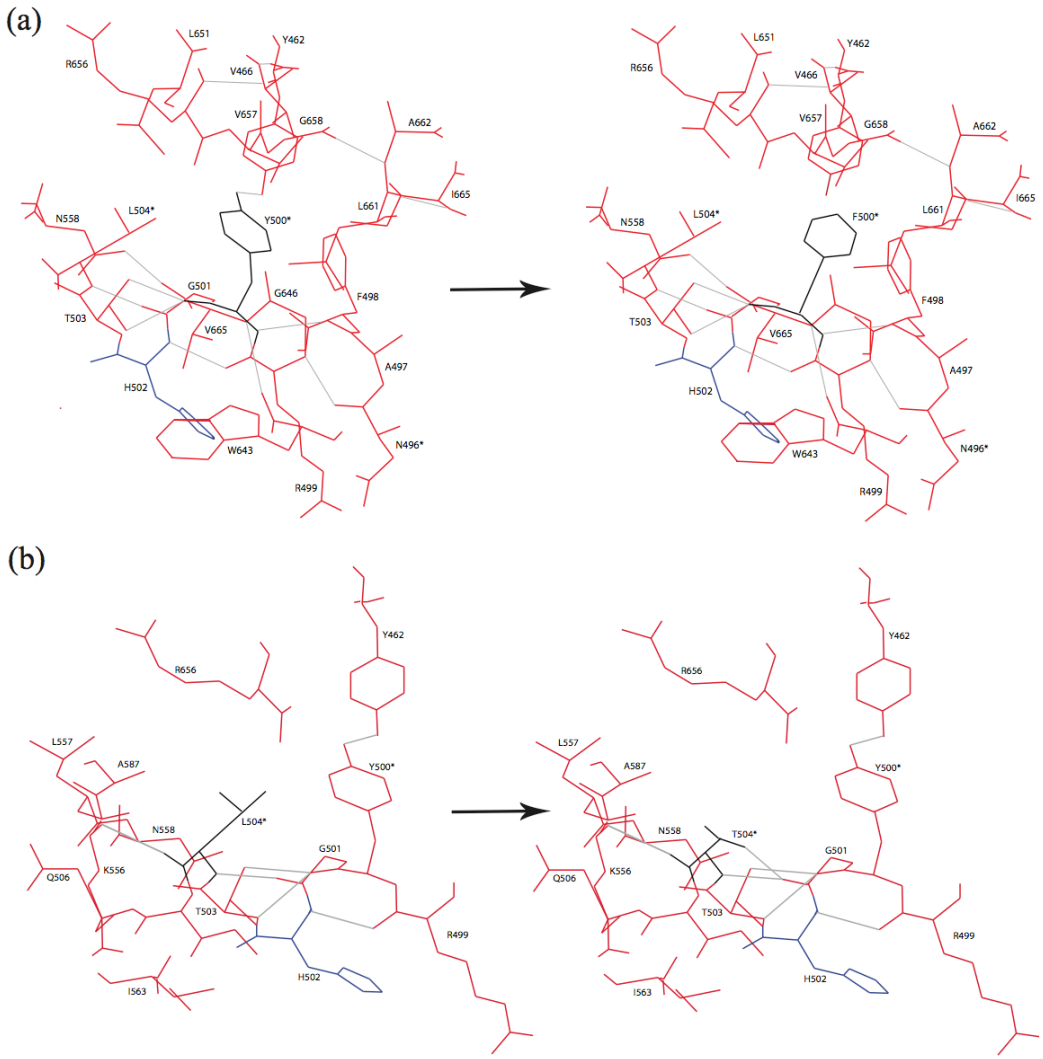


Figure 3.4. Effect of the Y500F and L504T mutation on hydrogen bonding within the myeloperoxidase structure. (a) Tyr/Phe 500 and (b) Leu/Thr 504 are seen in black, the heme ligand His 502 in blue, and hydrogen bonding in grey. An "*" denotes a positively selected site identified in Chapter 2 (N496, Y500, L504).

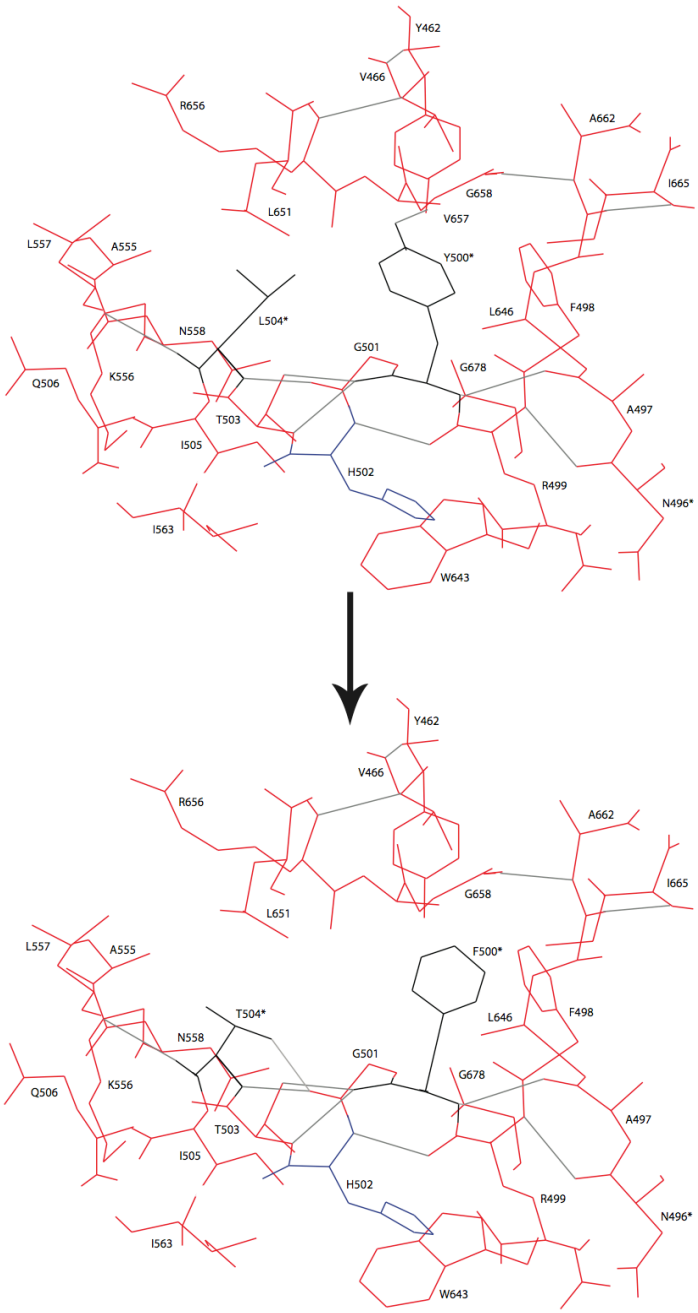


Figure 3.5. Effect of the double mutation, Y500F-L504T, on hydrogen bonding within the myeloperoxidase structure. Tyr 500-Leu504 and Phe 500-Thr 504 are seen in black, the heme ligand His 502 in blue, and hydrogen bonding in grey. An "*" denotes a positively selected site identified in Chapter 2 (N496, Y500, L504).

Other non-covalent interactions and steric hindrances (not assessed) may be disrupted as a consequence of these mutations and, in turn, may further compromise the structural integrity of MPO. To test the validity of our predictions as to the impact of these mutations on the structure of MPO, biosynthesis and enzymatic activity of mutant proteins expressed in a heterologous system was investigated.

3.3.2 *In vitro* site-directed mutagenesis

Site-directed mutagenesis was performed using pcDNA-MPO as template DNA (Table 3.5 in Section 3.2.1) with the aim of introducing five independent mutations in the MPO coding sequence at sites that are under positive selection (identified in Chapter 2). One is in the propeptide region (R80M) and four are in the heme binding pocket of the enzyme (N496F, Y500F, L504T, and the double mutant Y500F-L504T). The putative impact on hydrogen bond formation of mutating these positively selected sites in the heme binding pocket to a more ancestral state was assessed in the previous section and in Chapter 2. Sequencing analyses revealed that the N496F mutation was not successful and that the template, pcDNA-MPO, contained four missense mutations. At amino acid level these were L15S, G55E, V73M and M341T. Restriction analysis and subsequent ligation of mutant plasmids and pMPO (Table 3.5 in Section 3.2.1) were utilised to remove these unwanted variants in each of our mutant plasmids. The R80M mutation was lost during restriction analysis due to its close proximity to three of these missense mutants. Mutagenesis to re-introduce this mutation, using pMPO as template DNA, was carried out but was not successful. Further analyses of biosynthesis and function were carried out on wild type (WT) MPO and our three successful mutants in the heme pocket, Y500F, L504T and Y500F-L504T.

3.3.3 Assessing the effect of mutating positively selected sites on the biosynthesis of MPO

To determine the impact of these mutations on the biosynthesis of MPO, pulse-chase analysis using HEK 293 cells transiently and stably expressing WT and mutant MPO (Y500F, L504T and Y500F-L504T (Double)) was performed. All cell lines were biosynthetically radiolabeled with [^{35}S]-methionine for 1 h and chased for 0 and 20 h prior to immunoprecipitation of cell and medium fractions with MPO antiserum. Results from stable cell lines mirrored those determined from transient lines. Therefore, the following are the findings from the stable cell lines.

Each of the mutant-expressing cell lines synthesized 90-kDa precursor and 75-kDa intermediate species of MPO (Hansson, Olsson, Nauseef 2006) after pulse-labeling (Figure 3.6). In cells expressing normal MPO, proteolytic processing into mature MPO occurred, represented by the appearance of the 59-kDa heavy subunit of mature MPO. In contrast to the fate of normal MPO, mutant MPO precursors were not efficiently processed into mature enzyme. In Figure 3.6b, variation in the normal MPO fractions at 0 h is noted, this is due to the short-lived nature of the 75-kDa intermediate (Hansson, Olsson, Nauseef 2006). To quantitate the overall fate of MPO precursors in stable transfectants during the chase period, the fraction of 90-kDa precursor and 59-kDa subunit present at 20 hours and the ratio of 90-kDa precursor to 59-kDa mature heavy subunit at the end of the chase period were calculated (Table 3.9). For cells expressing normal MPO, this ratio was 0.74 ± 0.17, respectively (n = 7). In contrast, the failure of mutant MPO precursors to be processed was best illustrated by the excess 90-kDa relative to mature MPO, with 90-kDa:59-kDa for each of three mutants more than two-fold that of normally processed MPO (Table 3.9). Taken together, these data suggest that Y500F and L504T, alone or together, resulted in defective processing of MPO precursors into mature subunits. In addition, WT and mutant MPO secreted similar levels of the 90 kDa protein. This indicates that normal amounts of the mutant 90 kDa species entered the secretory pathway despite a lack of efficient processing of the mutant

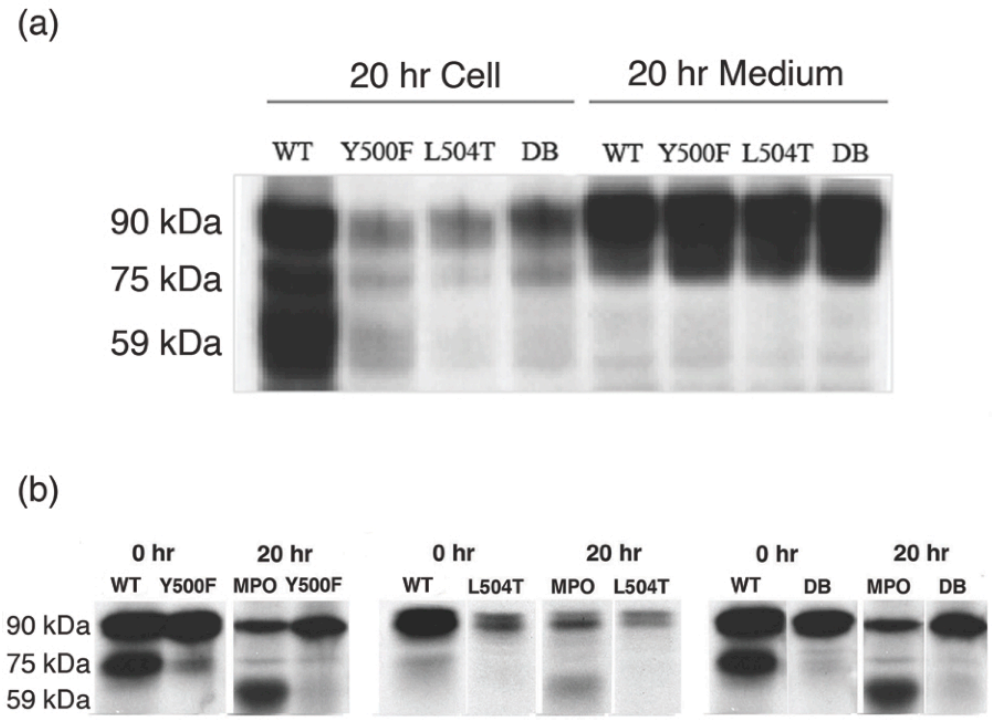


Figure 3.6. Biosynthesis of wild type (WT) and mutant MPO. HEK cells stably expressing wild type (WT) or mutant MPO (Y500F, L504T, Y500F-L504T) were pulse-labeled with ^{35}S and chased at 0 and 20 h intervals. Cell lysates at 20 h in the cell fraction and culture medium were collected and MPO-related protein was immunoprecipitated. Immunoprecipitated were analysed by SDS-PAGE and autoradiography. The 90 kDa species in both the cell and media fractions consist of both the inactive heme-free apoproMPO and enzymatically active proMPO precursors. The propeptide region is cleaved, generating the short-lived 75 kDa intermediate which undergoes further processing to yield mature MPO (depicted by the 59 kDa species). Any 90 kDa species that fails to undergo propeptide cleavage is constitutively secreted ((a) 20 hr medium fraction). This image was generated using (a) transient cell lines (20 hr cell and medium fraction) and (b) stable cell lines (0 and 20 hr cell fractions) expressing normal and mutant MPO.

Table 3.9: Cellular MPO-related protein.

Cell type	90kDa:59kDa
Normal	0.74 ± 0.17 (n=7)
Y500F	2.06 ± 0.36 (n=3)
L504T	1.65 ± 0.39 (n=4)
Double	3.22 ± 0.56 (n=3)

Ratio of the densitometric calculations of percentage MPO-related protein in cell fractions following pulse-chase analysis at 20 h (Mean ± SEM).

MPO (Figure 3.6). Since heme incorporation by apoproMPO to form proMPO is a prerequisite for normal proteolytic processing and maturation of MPO, it is reasoned here that the defective processing of Y500F and L504T, separately and together, may compromise the activity of the mutant protein products.

3.3.4 Effect of mutating positively selected sites on peroxidation and chlorination activity

To assess the impact of Y500F and L504T on the function of MPO, both peroxidase and chlorination activity of lysates from cells stably expressing mutant MPO were measured. Peroxidation is detected using TMB, which is oxidised, yielding a blue coloured cation free radical which can be detected spectrophotometrically. Chlorination is detected using the nonfluorescent APF substrate, which is cleaved by hypochlorite to generate fluorescin. Enzymatic activities were normalised to the levels of MPO-related proteins in each mutant, as judged by immunoblotting and subsequent densitometry (Figure 3.7a). Beta actin was utilised as a loading control in immunoblotting (Figure 3.7b). It was found that the percentage of MPO-related protein in mutant samples, Y500F, L504T and Y500F-L504T, relative to WT MPO (100%) were 52.94 ± 1.54, 21.59 ± 1.24 and 34.08 ± 1.50 % respectively (n = 4). To calculate relative specific activity, the percentage peroxidation and chlorination activity with respect to WT MPO was normalized to the amount of MPO-related protein in each cell lysate; see Table 3.10.

There was a significant reduction in the peroxidase activity of each mutant relative to that of WT MPO (control) (n = 4) (Figure 3.8a and Table 3.10). The relative specific activity of Y500F was ~ 72 % that of WT MPO ($p = 0.017$). There was a significant ($p < 0.0001$) drop in activity for the L504T and the double mutant, Y500F-L504T, relative to WT MPO, with ~ 1.1 and ~ 21 % relative specific activity, respectively. The combination mutant, Y500F-L504T, also undergoes a significant loss in peroxidases activity; however, it shows a slight increase in activity over that of the single L504T mutant, suggesting a compensatory evolutionary affect. These reductions in activity were far greater than the reduction in cell-associated MPO protein for the corresponding

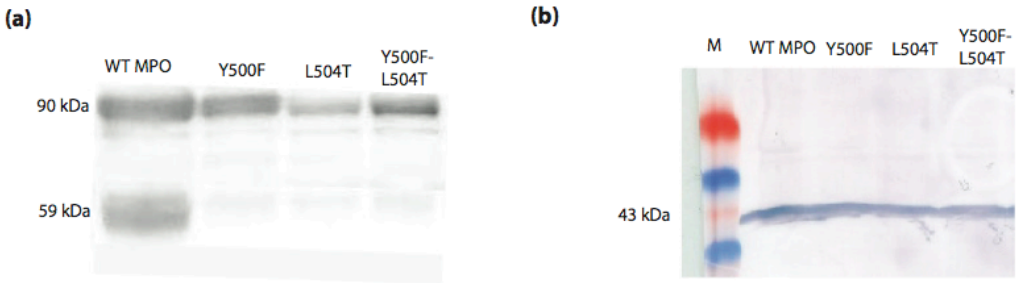


Figure 3.7. Immunoblotting of MPO-related protein and β-actin loading control. (a) MPO-related protein (90- and 59-kDa protein) immunoblot and (b) β-actin loading control immunblot (43-kDa protein).

Table 3.10: Relative Specific Activity.

	Activity (%)	
	Peroxidation	**Chlorination**
MPO	100	100
Y500F	72.39 ± 5.75*	0.00 ± 0
L504T	01.11 ± 1.11⌠	0.00 ± 0
Y500F-L504T	20.71 ± 2.31┌	0.00 ± 0

Percentage peroxidation and chlorination with respect to wild type MPO, normalized to amount of MPO-related protein in cell lysate (Mean ± SEM, *n* = 4). The significance values (paired two-tailed *t*-test) are as follows: *: $p = 1.72 \times 10^{-2}$, ⌠: $p = 3.15 \times 10^{-6}$, and ┌: $p = 5.44 \times 10^{-5}$.

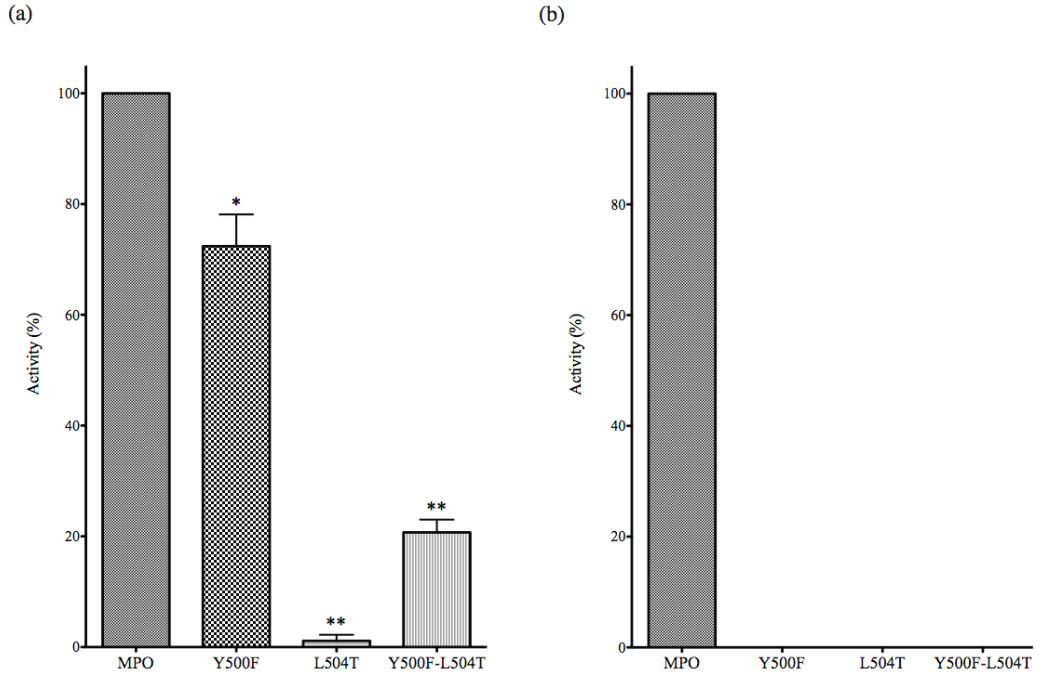


Figure 3.8. Myeloperoxidase activity. Percentage (a) peroxidation and (b) chlorination activity with respect to wild type MPO, normalized to amount of MPO-related protein in cell lysate. * $p < 0.02$, ** $p < 0.0001$.

mutants, as judged from the biosynthesis analyses. Given that very little of the 59-kDa mature MPO was detected in the pulse-chase analyses of the mutants, the activity detected here must reflect the contribution of the proMPO precursor. Thus, the low levels of specific activity of the mutant 90-kDa mutant products suggest that a large fraction of the 90-kDa protein synthesized existed as the heme-free proMPO with profoundly defective activity. It appears that these mutations in the heme pocket did not completely inhibit the incorporation of heme, as the mutants exhibited peroxidase activity, albeit significantly reduced ($p < 0.01$). However, these mutations completely ablated the capacity of the enzyme to produce HOCl, where correct incorporation of the heme co-factor is essential (Figure 3.8b and Table 3.10).

Overall, these data point to a functional effect for Y500 and L504, identified as being under positive selection using *in silico* methods and evolutionary theory. The analyses show that these residues are essential for stable acquisition of heme and subsequent proteolytic processing of MPO precursors, and for normal catalytic activity.

3.4 Discussion

Positive Darwinian selection is the process by which beneficial mutations in a population are retained and fixed, and is considered synonymous with protein functional shift. In general, one of the resultant gene copies that arises following a gene duplication event has increased freedom to explore mutational space, while the other copy executes the original function of the gene. Mutations that prove beneficial are retained in this new copy of the original gene through the process of positive selection, and over time this pressure can give rise to new functions, a process known as neofunctionalisation (Hughes 1999). Conserved positions and/or those mutations that do not confer a functional advantage may be under the influence of purifying/negative selection and/or as a result of genetic drift (selectively neutral). Therefore, it follows that functional shift or protein diversification and protein specialisation within multigene families is driven to a large extent by positive selection (Levasseur et al. 2006a). In chapter 2 of this thesis the evolutionary relationship of the medically important MHP family of enzymes has been fully resolved. Specific amino acid residues are seen to be responsible for the diversification of enzyme function in this family. MPO is unique among members of the functionally diverse MHP family of enzymes because of its capacity to oxidize chloride at physiologic pH and generate HOCl. This study focused on human MPO and investigated the functional effect of mutating Y500 and L504, residues predicted *in silico* to be under positive selection, on the synthesis and function of MPO *in vitro*.

Human MPO is encoded by a single gene located on the long arm of chromosome 17 and its biosynthesis, although not fully characterized, has had many of its features elucidated. Critical for the proper structure and function is the acquisition of heme by apoproMPO in the endoplasmic reticulum, resulting in generation of enzymatically active proMPO. The heme group in all members of the MHP family is covalently bound to the protein backbone by ester bonds with conserved aspartate and glutamate residues. In addition, MPO has a third covalent bond, a sulfonium linkage between the heme and M409. Covalent bond formation results from an autooxidation event in the ER and is believed to protect the vinyl groups of the heme from oxidation during the enzyme-

catalysed generation of highly acitve hypohalous acids (Colas, Ortiz de Montellano 2003; Huang, Wojciechowski, Ortiz de Montellano 2006). Heme acquisition is a prerequisite for the proteolytic processing of proMPO into mature MPO, as inhibition of heme synthesis causes an arrest of MPO biosynthesis at the apoproMPO stage. The clinical relevance of heme acquisition and proper processing of MPO is illustrated by several genotypes of inherited MPO deficiency. Patients with inherited deficiency due to R499C, G501S and R569W (Nauseef, Cogley, McCormick 1996; Goedken et al. 2007) missense mutations in MPO have a maturational arrest in MPO biosynthesis at the apoproMPO stage; in the absence of heme incorporation into apoproMPO, neither proMPO nor enzymatically active MPO is formed, resulting in peroxidase-deficient neutrophils.

The impact of the mutations at R499 and G501, both near the proximal heme ligand at H502 in MPO, is especially pertinent to the two residues studied in this chapter. The *in silico* structural analyses have revealed that Y500F (loss of an hydroxyl group) and L504T (gain of an hydroxyl group) would have potential implications for the structure of the protein by perturbing the hydrogen bonding around the heme binding pocket. Although many studies have hypothesized that nucleotide divergence is the driving force for neofunctionalization, very few have investigated the link between positive selection and functional divergence experimentally. Levasseur *et al.* (2006) and Yokoyama *et al.* (2008) investigated the role that positive selection plays in functional diversification but obtained contrasting results. Levasseur *et al.* (2006) studied the fungal lipase/feruloyl esterase A family for signatures of positive selection. Following *in vitro* site directed mutatgenesis of identified positions, their results clearly demonstrated that certain amino acids under positive selection were involved in the functional shift. In contrast, Yokoyama *et al.* (2008) investigated the evolution of phenotypic adaptations of visual pigments in vertebrates. They too performed *in vitro* mutational analyses of sites under positive selection. However, their findings revealed no significant influence of positive selection on the adaptation of rhodopsin sensitivity. These studies highlight the necessity to provide experimental evidence to support/validate computational analyses.

To test the hypothesis that positive selection is a driving force in the evolutionary diversification of the MHP, the impact of mutations Y500F, L504T and Y500F-L504T on MPO have been examined at a molecular and phenotypic level. The pulse-chase and functional analyses of stably transfected cell lines expressing mutant forms of MPO revealed a profound effect of the mutations on the cellular fate and activity of MPO. Although the biosynthesis of a 90-kDa MPO precursor proceeded normally in transfectants expressing mutant MPO, subsequent proteolytic processing was impaired in all three mutants. Failure to generate mature enzyme from precursor was most profound in the double mutant, where the 90-kDa: 59-kDa ratio, an indication of efficient proteolytic processing, was > 4-fold greater in Y500F-L504T in comparison to that seen in normal MPO. Given that formation of proMPO is a prerequisite for generation of mature MPO and that heme acquisition by apoproMPO results in proMPO, it was reasoned that mutations at Y500 and L504 compromised stable heme binding by mutant apoproMPO. In fact, cell lysates from transfectants expressing mutant MPO exhibited depressed peroxidase activity. It is noteworthy that L504T impaired peroxidase activity much more dramatically than did Y500F, whereas there was relatively more 90-kDa MPO-related protein at 20 hours in L504T- than in Y500F-expressing cells. If L504T resulted in a more stable 90-kDa apoproMPO, the amount of active MPO-related protein in cell lysates would be less and, consequently, the calculated relative specific activity of L504T lower. None of the mutants supported chlorination, consistent with significant disruption of the integrity of the heme pocket by the mutations.

Taken together, these data support the concept that Y500 and L504, positively selected sites within the MPO protein, are involved in the observed protein functional shift. The impact of the present mutations on the peroxidation and chlorination activity of MPO revealed the biological significance of the *in silico* predictions, as the unique property of MPO to produce the potent oxidant HOCl was eliminated following mutation of positively selected residues. This loss of function suggests that these residues have been positively selected in the MPO lineage, as they accommodated the beneficial new function of chlorination activity and the production of HOCl. Like the molecular

phenotypes of R499 and G501, mutations identified in inherited MPO deficiency and predicted to disrupt the environs of the proximal heme pocket of MPO, the mutants in this study resulted in aberrant processing of MPO precursors and loss of functional mature enzyme. Further studies on the biosynthesis of mutant MPO versus normal MPO could be performed with the aim of pinpointing at what stage maturational arrest occurs in the biosynthesis of mutant MPO, by co-immunoprecipitation of MPO associated with the molecular chaperones involved in heme acquisition, CRT and CLN (DeLeo et al. 1998; Goedken et al. 2007). Heme acquisition may be further assessed by radiolabeling with δ-[^{14}C]aminolevulinic acid, which is a precursor in heme synthesis (DeLeo et al. 1998; Goedken et al. 2007). Determining the Soret band of each mutant may indicate if the integrity of the heme pocket is disrupted upon mutation (Goedken et al. 2007).

The results of this chapter have shown that positive selection signifies a functional shift in the MHP multigene family of enzymes, illustrating how evolutionary predictions can successfully identify residues essential for unique protein functions. Use of evolutionary biology as a predictive tool for targeted mutagenesis has a major role to play in the future elucidation of protein biology and evolutionary medicine.

In this chapter the site directed mutagenesis was guided by the phylogenetic tree, i.e. the mutations generated were to the ancestral state. It is possible, therefore, not only to do site directed mutagenesis, but, entire ancestral gene genesis, thereby, generating an ancient protein. The generation of an ancient protein and its characterisation is the focus of the final results chapter (chapter 4).

Chapter 4

Resurrection and preliminary biochemical characterisation of an ancient plant heme peroxidase (~ 113 million years old).

4.1 Introduction

The evolution of the animal heme peroxidase has been the focal point of the previous two chapters. This chapter shifts focus to the other major heme peroxidase superfamily, the plant heme peroxidases, that have arisen independently but which also have heme dependency and peroxidase activity. Like their animal counterparts, plant heme peroxidases are also oxidoreductases. Plant peroxidases are involved in wound healing, lignification and play a role in regulation of germination (Azevedo et al. 2003; dos Santos et al. 2004). Horseradish peroxidase (HRP) and soyabean peroxidase (SBP) are used extensively in the biopharmaceutical and biotechnology sectors. Both have major applications in biosensor and immunoassay technologies. These peroxidases also have important features in organic chemistry, bioremediation and therapeutics (Ryan, Carolan, O'Fagain 2006). Their operational potential/efficiency in these sectors is limited by the thermal and oxidative (hydrogen peroxide (H_2O_2) tolerance) stabilities of the enzymes.

HRP A2, HRP C and SBP are all members of Class III secretory plant peroxidases. The evolutionary relationship of the Class III subgroup of plant peroxidases has been fully resolved (Duroux, Welinder 2003). Although these plant peroxidases are closely related, their enzymatic stabilites are considerably different. Previous studies have shown that SBP is more thermostable than HRP A2 and HRP C , with HRP A2 being the least thermally stable of the three (McEldoon, Dordick 1996; Kamal, Behere 2003). These three Class III peroxidases also exhibit varying stability towards their primary substrate, H_2O_2, and an apparent inverse pattern between this substrate tolerance and thermal stability. In contrast SBP displays an increased thermal stability, and very poor tolerance to H_2O_2 (McEldoon, Dordick 1996; Henriksen et al. 2001).

The relationships between the extant plant peroxidases of interest are shown in Figure 4.1a. Their thermal stabilities and H_2O_2 tolerances are given beside their names. From Figure 4.1a, it is clear that the ancestral sequence (ancestral reconstruction) on the tree is the ancestral node of all of these properties of interest. As such, one can hypothesise that

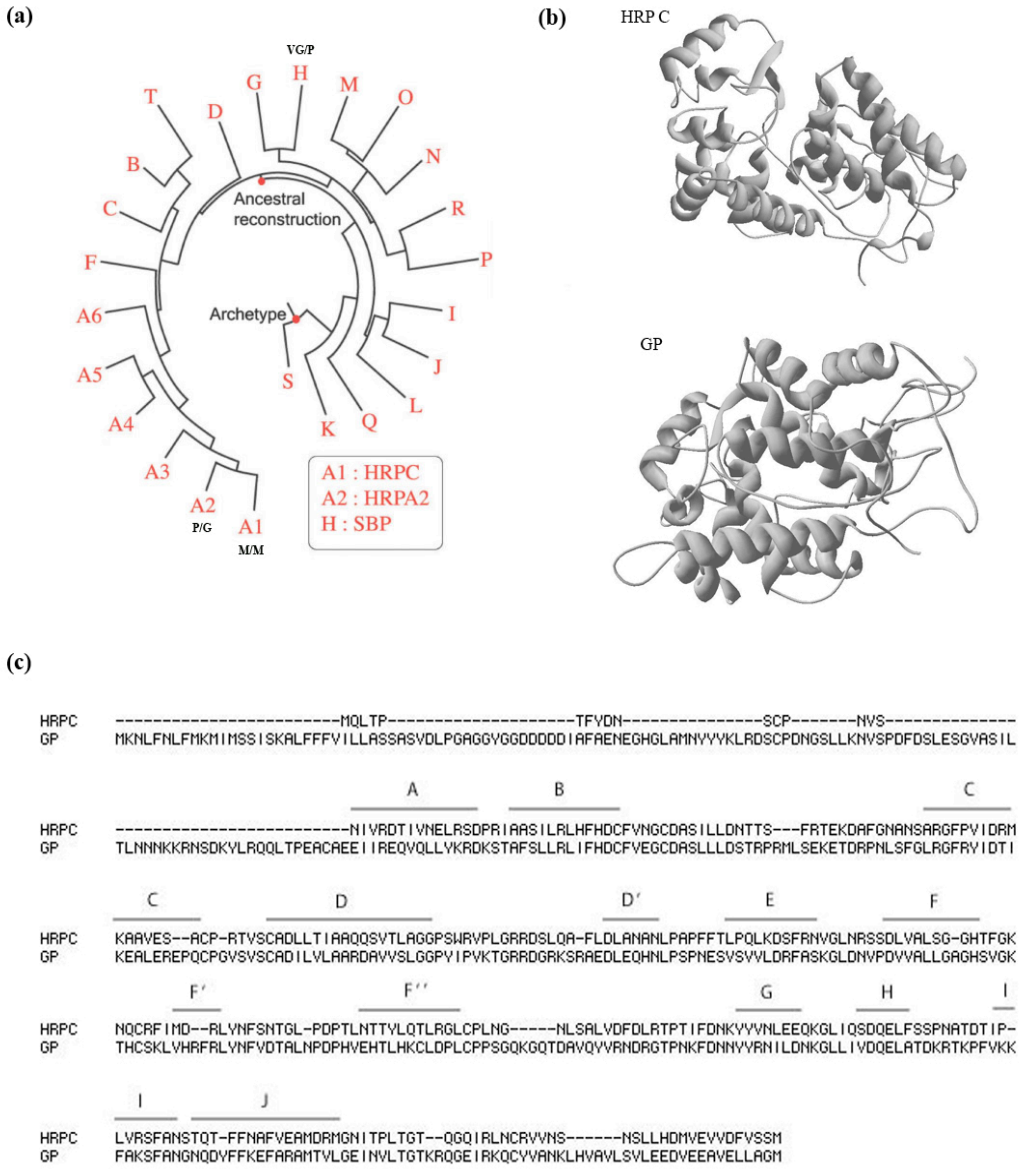


Figure 4.1. Ancestral plant heme peroxidase location on phylogenetic tree, amino acid sequence and 3-D structure from homology modeling. (a) The location of the resurrected ancestral plant peroxidase (ancestral reconstruction) on the reduced phylogeny of Class III plant peroxidases, adapted from Ryan, O'Connell and Ó Fágáin 2008 and Duroux and Welinder 2003. Stabilities at the labels of interest are in the denoted in the following order thermal/oxidative, where VG: very good, G: good, M: moderate, P: poor. (b) 3-D structure of extant HRP C PDB code 1W4Y (*top*) and homology modeled 3-D structure of the resurrected archetypal plant peroxidase (GP)

based on chain A of *Arabidopsis thaliana* A2, PDB code 1pa2 (*bottom*). The ancient enzyme contains more unstructured loops than its extant counterpart. Only 40 % sequence similarity existed between the archetypal protein sequence and the selected best-fit template sequence during homology modeling. Therefore, the first 100 amino acids of the GP are not modeled. (c) Alignment of extant HRP C (PDB code 1W4Y) and reconstructed ancestral plant heme peroxidase (GP). Helices are highlighted and labeled based on literature (Gajhede et al. 1997; Ryan, O'Connell, O'Fagain 2008).

it may hold a desirable mixture/complement of these characteristics of interest. Recreating ancient proteins (paleomolecular biochemistry) can provide information on the functional diversity of multigene families and even resurrect favourable biochemical properties that may have been lost over time (Chang 2003; Thornton *et al.* 2003).

In a previous study within the group (Ryan, O'Connell, O'Fagain 2008), a reduced phylogeny of the Class III plant peroxidases was generated and using this representative phylogeny, ML approach was implemented to reconstruct all ancestral nodes on the phylogeny, including the most recent common ancestor (MRCA) of HRP A2, HRP C and SBP: see Figure 4.1a for location (ancestral reconstruction) on the reduced phylogeny. Resurrecting this ancient plant peroxidase will allow for the direct study of the molecular evolution of these extant peroxidases

This ancestral sequence existed in the ancestor of HRP A2, HRP C and SBP. From the botany literature it is known that all species sharing this common ancestor diverged from the core eudicots into rosids (108-109 million years old (MYO)) and asterids/caryphyllales (105 ± 1 MYO). These records suggest that this ancient sequence is approximately 113 MYO (± 1 MY), placing it in the Cretaceous period of the Mesozoic era (Orndorff et al. 2009).

This chapter describes (i) the successful expression of an active 113 million year old plant heme peroxidase *de novo,* and (ii) the assessment of its thermal and oxidative stabilities. With respect to its extant counterparts, this ancient protein exhibits a moderate thermal stability with an increased tolerance to H_2O_2, which may be favourable in many industrial applications. Further biochemical characterisations will provide greater insights into the evolution of these plant peroxidases and highlight the operational potential in the biopharmaceutical and biotechnology sectors of this ancestral enzyme.

4.2 Methodology

4.2.1 Biological Materials

The following tables detail the biological materials used in this study:

Table 4.1: Bacterial strains used in this study.

Bacterial Strain	Genotype	Source
E. coli XL10-Gold	Tet^r D(mcrA)183 D(mcrCB-hsdSMR-mrr)173 endA1 supE44 thi-1 recA1 gyrA96 relA1 lac Hte F'[proAB lacI ^{q}ZDM15 Tn10 (Tet^r) Amy Cam^r]	Stratagene
E. coli XL1-Blue	recA1, endA1, gyrA96, thi-1, hsdR17(r_K^-,m_K^+), supE44, relA1, λ^-, lac^-, [F' proAB, $lacI^q$ZΔ M15, Tn10(Tet^r)].	Stratagene
E. coli JM109	endA1 glnV44 thi-1 relA1 gyrA96 recA1 $mcrB^+$ Δ(lac-proAB) e14- [F' traD36 $proAB^+$ $lacI^q$ lacZΔM15] hsdR17($r_K^-m_K^+$)	Stratagene

Table 4.2: Plasmids used in this study.

Plasmid	Description	Source
pGSLink	High-level expression vector for expression of N- or C-terminal 6xHis-tagged fusion proteins linked to protein of interest via a flexible peptide linker.	Loughran *et al.* 2006
pBR_I	pQE_pelB_HRP_His vector with wildtype HRP gene present.	Dr. Barry Ryan
pGA	GENEART transport vector harbouring the archetypal gene sequence (GP)	This study
pGS-pelB	pGSLink with pelB leader sequence inserted via *Nco* I-*Not* I-*BamH* I cloning.	This study
pGP	pGS-pelB with GP inserted via *Not* I-*Bgl* II cloning	This study
pUC18	Transformation control vector	Stratagene

Table 4.3: Oligonucleotides used in this study.

Target	Primer Sequence
pelB For	5'-CATGCCATGGGCATGAAATACCTGCTGCCG-3'
pelB Rev	5'-CGGGATCCGCGGCCGCGGCCATCGCCGGCTGGG-3'
GP For	5'-AAGGAAAAAAGCGGCCGCCATGAAAAACCTGTTTAA-3'
GP Rev	5'-GGAAGATCTCATACCTGCCAGCAGTTC-3'
pQE For*	5'-GTATCACGAGGCCCTTTCGTCT-3'
pQE Rev*	5'-CATTACTGGATCTATCAACAGGAG-3'

*Sequencing primers

4.2.2 DNA Manipulation

Methodology for plasmid preparation, restriction digestion of DNA, ligation of DNA, agarose gel electrophoresis and DNA quantification and sequencing can be found in Section 3.2.2. All medium components, supplements and complex buffers are given in the Appendix.

4.2.2.1 PCR product clean-up

Amplified PCR products were cleaned for further analysis using Bioline's DNAce Quick-Clean as per manufacturer's protocol. Briefly, an equal volume (2 volumes for pelB amplified DNA) of DNAce Quick-Clean was added to amplified GP sample and mixed by pipetting. The solution was incubated at room temperature for 5 min (8 min for pelB sample). Samples were centrifuged for 10 min at 10,000 x *g*; supernatant was discarded. Pellets were washed in 50 μL of 70 % (v/v) ethanol and dried under heat-vacuum for 5 min. Pellets were resuspended in 50 μL sterile H_2O (sH_2O) and used either immediately or stored at 4 ºC.

4.2.2.2 Dephosphorylation of linearised plasmid DNA

To prevent plasmid recircularisation or plasmid-plasmid ligation, treatment with calf intestinal phosphatase (CIP) (0.5 U/μg) (New England Biosciences) is required whereby the 5' phosphates on the linearised plasmid are removed. Digested plasmid DNA (<100 ng/μL) was treated in a total reaction volume of 100 μL; linearised plasmid: 30 μL, CIP: 5 μL, CIP buffer: 7 μL, sH_2O: 58 μL, and incubated at 37 ºC for 1 h followed by heating to 75 ºC for 10 min to denature the enzyme. CIP-treated DNA was stored at 4 ºC/-20 ºC until required.

4.2.2.3 Gel DNA fragment extraction

DNA fragments were extracted from agarose gel using the HiYield™ Gel/PCR DNA Extraction Kit as per manufacturer's protocol. Briefly, DNA fragments were visualised under UV light and a gel slice containing DNA was excised (~300 mg) and transferred

to a microcentrifuge tube. DF buffer (500 μL) was added and the sample was mixed by vortexing and incubated at 55 ºC for 15 min, inverting the tube every 3 min, to dissolve the gel slice. The sample was then placed into the DF column collection tube and centrifuged for 30 sec at 10,000 x *g*; flow-through was discarded. The DF column was washed with 500 μL Wash Buffer by centrifugation as before; flow-through was discarded. The DNA-bound column was dried by centrifugation for at 10,000 x *g* for 2 min. The dried DNA-bound column was transferred to a fresh microcentrifuge tube and Elution Buffer (30 μL) was applied to the centre of the column matrix and allowed to stand for 2 min before elution of DNA by centrifugation (10,000 x *g*: 2 min). Purified DNA was stored at 4 ºC/-20 ºC until required.

4.2.2.4 Preparation of competent cells

A glycerol stock was streaked on Luria-Bertani (LB) agar and incubated overnight at 37 ºC. A single colony was used to inoculate 10 mL of LB medium which was then incubated overnight at 37 ºC and 220 rpm (TS NETwise Gallenkamp ORBI-SAFE incubator). Overnight culture (1 mL) was added to 100 mL of pre-warmed LB medium in a 500 ml flask and the flask was then incubated at 37 ºC and 220 rpm (TS NETwise Gallenkamp ORBI-SAFE incubator) until an OD_{600} 0.5-0.6 was reached. The culture was centrifuged at 4,500 x *g* for 5 min at 4ºC. The cell pellet was gently resuspended in 30 mL of ice cold TFB1 buffer and incubated on ice for 10 min. The solution was then centrifuged as before and the cell pellet was gently resuspended in 4 ml of ice cold TFB2 buffer and incubated on ice for 15-60 min. The cell suspension was aliquoted into appropriate volumes, flash frozen and stored at -80ºC.

4.2.2.5 Transformation

A microfuge tube of competent cells was allowed to thaw on ice and the cells were transformed by heat shock. Plasmid DNA (2 μL; < 100 ng/μL) was added to 200 μL of the competent cells. The transformation reaction was mixed gently and then incubated on ice for 30 min. The cells were then heat shocked for 45 sec at 42 ºC and then placed

on ice for a further 2 min. Super optimal broth with catabolite repression (SOC) medium (0.8 mL) was added, this was followed by incubation at 37 ºC for 45 min with shaking at 150 rpm (TS NETwise Gallenkamp ORBI-SAFE incubator). This resulting transformation mixture (100 μL) was plated on appropriate selective LB agar and incubated at 37 ºC overnight. Resultant experimental colonies were used to prepare broth cultures for further analysis.

4.2.2.6 Determination of competent cell efficiency

Competent cell efficiency was defined as the number of colony forming units per μg of transformed plasmid DNA. A stock of pUC18 plasmid (25 ng/μL) (Table 4.2) was diluted to 250 pg/μL, 25 pg/μL and 2.5 pg/μL. An aliquot (2 μL) of each dilution was transformed independently (Section 4.2.2.5) and the transformation efficiency was determined from the number of colonies formed.

4.2.2.7 Two-step cloning

A two-step cloning strategy was utilised to generate an expression plasmid bearing the pelB leader sequence and the archetypal gene sequence (GP). Primers were designed for independent amplification of the pelB leader sequence from the pBR_I vector (*Nco* I and *Not* I-*Bam*H I restriction sites incorporated) and the GP sequence from the GENEART transport vector (pGA; *Not* I and *Bgl* II restriction sites incorporated); see Tables 4.2 and 4.3 for descriptions of plasmids and primers respectively.

The following PCR reaction mix and programme was utilised for DNA amplification:

Table 4.4: PCR reaction mix.

Component	Volume
sH_2O	38.5 µl
10X *Red Taq* reaction buffer	5 µl
Template DNA (< 500 ng)	1 µl
Forward primer (100 pmole)	1 µl
Reverse primer (100 pmole)	1 µl
dNTP mix (10 mM)	1 µl
Red Taq polymerase (5 U/µL)	2.5 µl

Table 4.5: PCR programme.

Segment	Cycles	Temperature	Time
1 (Denaturation)	1	95 ºC	2 min 30 sec
2 (Cycling: Denaturation-Annealing-Elongation)	18	95 ºC	50 sec
		T_{anneal} (ºC)	50 sec
		72 ºC	1 min/kb
3 (Elongation)	1	72 ºC	5 min

where T_{anneal} = [2x (A/T content) + (4x (G/C content)] – 5 °C

All PCR reactions were performed in a Thermo Electron Corporation PX2 Thermal Cycler. Amplified PCR products (pelB and GP) were cleaned using the commercial DNAce Quick Clean® (Bioline) method (Section 4.2.2.1). The cleaned PCR products were then digested with the appropriate restriction enzymes.

The pGSLink vector (Table 4.2) was digested with *Nco* I and *Bam*H I restriction enzymes and, to prevent plasmid recircularisation or plasmid-plasmid ligation. The digested vector was treated with CIP to remove 5' phosphates (Section 4.2.2.2).

The restricted pelB PCR product and CIP treated vector digest were purified from agarose gels using the HiYield™ Gel/PCR DNA Extraction (Section 4.2.2.3). The gel purified restricted insert and vector were ligated overnight (Section 3.2.2.3).

The overnight ligation was transformed into *E. coli* XL10Gold and the resultant plasmid (pGS-pelB) was prepared (Sections 4.2.2.5 and 3.2.2.1 respectively). The pGS-pelB vector was then digested with *Not* I and *Bgl* II restriction enzymes and CIP treated to prevent plasmid recircularisation. The restricted GP PCR product and CIP treated vector digest were purified from agarose gel, ligated, transformed and the plasmid DNA (pGP) was purified as above and sequenced by Eurofins/MWG-Biotech, London (Section 3.2.2.7).

4.2.3 *In Silico* Protein Analysis

4.2.3.1 Homology modelling

Homology modeling of the archetypal plant peroxidase was performed as described in section 2.2.3.3. The optimum template selected for modeling was chain A of *Arabidopsis thaliana* A2, PDB code 1pa2. Caution noted: 40 % sequence similarity between the archetypal protein sequence and selected best-fit template sequence.

4.2.4 *In Vitro* Protein Analysis

4.2.4.1 Protein expression

An LB plate with the appropriate antibiotic was streaked with a stock of the strain containing the expression plasmid of interest. A single colony was selected to inoculate

10 mL of LB broth containing the appropriate antibiotic, and grown overnight at 37 °C and 220 rpm (TS NETwise Gallenkamp ORBI-SAFE incubator). A 250 mL conical flask comprising 100 mL of LB broth containing the appropriate antibiotic, 2 % w/v glucose and 2 mM $CaCl_2$ was inoculated with 1 mL of the overnight culture. The culture was incubated at 37 °C and 220 rpm (TS NETwise Gallenkamp ORBI-SAFE incubator) until an $OD_{600nm} \approx 0.6$ was reached. The culture was centrifuged at 10,000 x *g* for 10 min at 4 °C. The supernatant was discarded and pellet was resuspended in the same volume of LB broth containing the appropriate antibiotic and 2 mM $CaCl_2$. Isopropyl β-D-1-thiogalactopyranoside (IPTG) was then added to a concentration of 50 μM to induce expression; a culture containing no IPTG was also expressed. The culture temperature was then reduced to 30 °C and it was incubated for a further 6 h. Cells were harvested by centrifugation at 4,700 x *g* for 5 min. The supernatant was stored at 4 °C until peroxidase activity analysis (Section 4.2.4.9) and cell pellet was stored at -20 °C until lysis.

4.2.4.2 Cell lysate preparation

Cell pellets from Section 4.2.4.1 were resuspended in 10 mL 50 mM sodium phosphate pH 7.5. The sample was sonicated on ice at 40 % amplitude for 90 sec with 6 sec pulses using a Branson Digital Sonifer®. Sonicated samples were centrifuged at 10,000 x *g* for 10 min at 4 ºC. The cleared lysate was filtered through a 0.45 μm sterile filter and stored at 4 °C until required for further analysis.

4.2.4.3 Optimisation of recombinant protein expression

Recombinant protein expression was performed as per Section 4.2.4.1, however, upon glucose depletion, cultures were induced with varying concentrations of IPTG (final concentrations: 0, 0.1, 0.5 and 1 mM). Cultures were incubated at 37 °C or 30 °C and 220 rpm (TS NETwise Gallenkamp ORBI-SAFE incubator). Samples were collected at 0, 2, 4 and 18 h time points. Cells were harvested by centrifugation at 4,700 x *g* for 5 min and the cell pellet was stored at -20 °C until lysis.

4.2.4.4 Optimum recombinant plant peroxidase expression

Protein expression was performed as per Section 4.2.4.1, however, upon glucose repletion, cultures were incubated at 30 °C and 220 rpm (TS NETwise Gallenkamp ORBI-SAFE incubator) for 18 h. Cells were harvested by centrifugation at 4,700 x *g* for 5 min and the cell pellet was stored at -20 °C until lysis.

4.2.4.5 Recombinant plant peroxidase purification by immobilized metal affinity chromatography (IMAC) using Ni-NTA resin

Ni-NTA resin was washed three times with 10 mL 50 mM sodium phosphate buffer pH7.5 to remove any traces of ethanol and to equilibrate the resin. Briefly, the wash step involved applying 10 mL of buffer to the resin in a sterile universal. The contents were swirled gently to mix and centrifuged at 3,000 x *g* for 10 min. The aqueous layer was removed without disturbing the resin. This was repeated three times. Clarified lysate (10 mL) was applied to the equilibrated resin and allowed to mix by tumbling at 4 °C for at least 1 h to allow binding to occur. The mixture was then poured into 1.5 x 20 cm column allowing the resin to settle. The flow through (FT) fraction was collected. The bound resin was then washed with 50 mM sodium phosphate pH 7.5 up to five times; wash fraction (W). The bound protein was then eluted with 50 mM sodium acetate pH 4.5 up to three times; elution fraction (E). Eluted fractions were dialysed overnight at 4 °C in 5 L of 50 mM sodium phosphate pH 7.5. All samples were subjected to SDS-PAGE. Purified eluted fractions were concentrated using Amicon Centricon® concentrations (Millipore) and filter sterilised and stored at 4 °C for further biochemical analysis.

4.2.4.6 Preparation of samples for SDS-PAGE analysis

An equal volume of 2X SDS loading buffer was added to the clarified lysate and post-purification samples. The mixture was boiled for 5 min and the cell debris was pelleted by centrifugation at 10,000 x *g* for 10 min and discarded. The samples were then subjected to SDS-PAGE as per Section 3.2.4.3 or stored at -20 ºC.

4.2.4.7 Coomassie Blue staining

The gel was placed in Coomassie stain for 30 minutes with gentle agitation. The gel was then placed in destain with constant agitation, and destain was changed 4 or 5 times at 1 h intervals until all background staining was removed from the gel. An image of the gel was then captured.

4.2.4.8 Protein quantification by bicinchoninic acid (BCA) assay

The bicinchoninic acid (BCA) described by Smith *et al*. (1985) was utilised to quantify total protein. A standard curve was created using bovine serum albumin (BSA) as the reference protein (0-1,000 μg/mL BSA concentrations). All experimental and standard samples (25 μL each) were added to 200 μL BCA Working Reagent (Sigma-Aldrich) and incubated in the dark at 37 °C for 30 min. Absorbance was read at 562 nm. A standard curve was created using BSA standards as the reference protein. Protein concentration of the unknown was determined from this standard curve.

4.2.4.9 Reinheitzahl number of recombinant archetypal plant peroxidase

The absorbance of purified recombinant archetypal plant peroxidase was recorded at 403 and 208 nm using quartz cuvettes. The Reinheitzahl number (RZ) was calculated using the formula:

$$RZ = {A_{403nm}}/{A_{280nm}} \qquad Eq.\ 4$$

4.2.4.10 Peroxidase activity assay (TMB assay)

3,3',5,5'-Tetramethylbenzidine dihydrochloride (TMB) (Sigma-Aldrich) was used as a reducing substrate to determine peroxidase activity of recombinant archetypal plant peroxidase (Ryan *et al.* 1994). TMB (1 mg) was dissolved in 200 μL of dimethylsulphoxide (DMSO) (Sigma-Aldrich). The mixture was then added to 9.8 mL

of 100 mM citric acid, pH5.5, and mixed thoroughly. 50µl of enzyme preparation (50 mM Sodium Phosphate, pH 7.5, as diluent) was added to a microtitre plate. Prior to commencement of the TMB assay, 4 µL of H_2O_2 (30% v/v) was added to the TMB/DMSO/citric acid mixture and mixed thoroughly. This mixture (150 µL) was immediately added to the 50µl enzyme preparations and incubated at room temperature. The absorbance at 620 nm was recorded after 7 min.

4.2.4.11 Preparation of extant commercial plant peroxidases

Commercial HRP A2 and HRP C were obtained from Biozyme and Sigma-Aldrich respectively. SBP was obtained from Quest International. Samples were prepared by dissolving in 50 mM sodium phosphate pH7.5 to a final concentration of 1 mg/mL. The crude SBP preparation was mixed for 20 min at room temperature and debris was pelleted by centrifugation at 4,000 x *g* for 15 min. Supernatant samples were further analysed.

4.2.4.12 Oxidative stability

The oxidative stability (H_2O_2 tolerance) of recombinant archetypal and extant commercial plant peroxidases was determined. In brief, enzyme preparations were exposed to varying concentrations of H_2O_2 (0, 5, 10, 15, 20, 40, 60 and 80 mM) for 30 min at 25 °C. H_2O_2 concentrations were determined spectrophotometrically at 240 nm using a molar extinction coefficient of 43.6 $M^{-1}cm^{-1}$ (Hernández-Ruiz *et al.* 2001). The residual activity of each sample was determined, where 0 mM H_2O_2 represented 100 % activity.

4.2.4.13 Thermal profile

A thermal profile of recombinant archetypal and extant commercial plant peroxidases was determined. Briefly, enzyme preparations were exposed to varying temperatures for 10 min (20, 30, 40, 50, 55, 60, 65, 70, 75 and 80 °C). The peroxidase activity of each sample was assessed as per Section 4.2.4.9. The residual activity of each sample was

determined, where activity at 20 °C represented 100 % activity. The temperature, at with 50 % of residual activity (T_{50}) was observed, was noted.

4.2.4.14 Thermal inactivation

For each enzyme preparation, thermal inactivation was performed at their respective T_{50}, as determined in Section 4.2.4.12. Samples were collected at 0, 1, 2, 4, 6, 8 and 10 min and their peroxidase activity was assessed at each of these time points as per Section 4.2.4.9. The residual activity of each sample was determined, where activity at 0 min represented 100 % activity. The half-life ($t_{1/2}$) was calculated using:

$$t_{1/2} = \ln 2 / k \qquad Eq.\ \ 5$$

where k is the rate constant. k was calculated using the *Enzfitter* programme (Biosoft, Cambridge, UK) by fitting the data to a single (first order) exponential decay model.

4.2.4.15 Recombinant archetypal plant peroxidase kinetics (ABTS assay)

2,2'-azino-bis(3-ethyl-benzthiazoline-6-sulphonic acid) (ABTS) was used as a reducing substrate to determine recombinant archetypal plant peroxidase kinetics. A 5 mM ABTS stock solution was prepared in ABTS buffer. From this stock, a range of ABTS standards were prepared (0 – 1 mM) in ABTS buffer. A 100 mM H_2O_2 stock solution was also prepared. In a microtitre plate, 2.5 μL 100 mM H_2O_2 and 222.5 μL of desired substrate concentrated were aliquoted. To initiate the reaction, 25 μL of enzyme preparation was added to each H_2O_2/ABTS mixture. The absorbance at 405 nm was recorded at 1 min intervals for a total time of 20 min. The change in absorbance per min (ΔA/min) was calculated for each substrate concentration.

4.3 Results

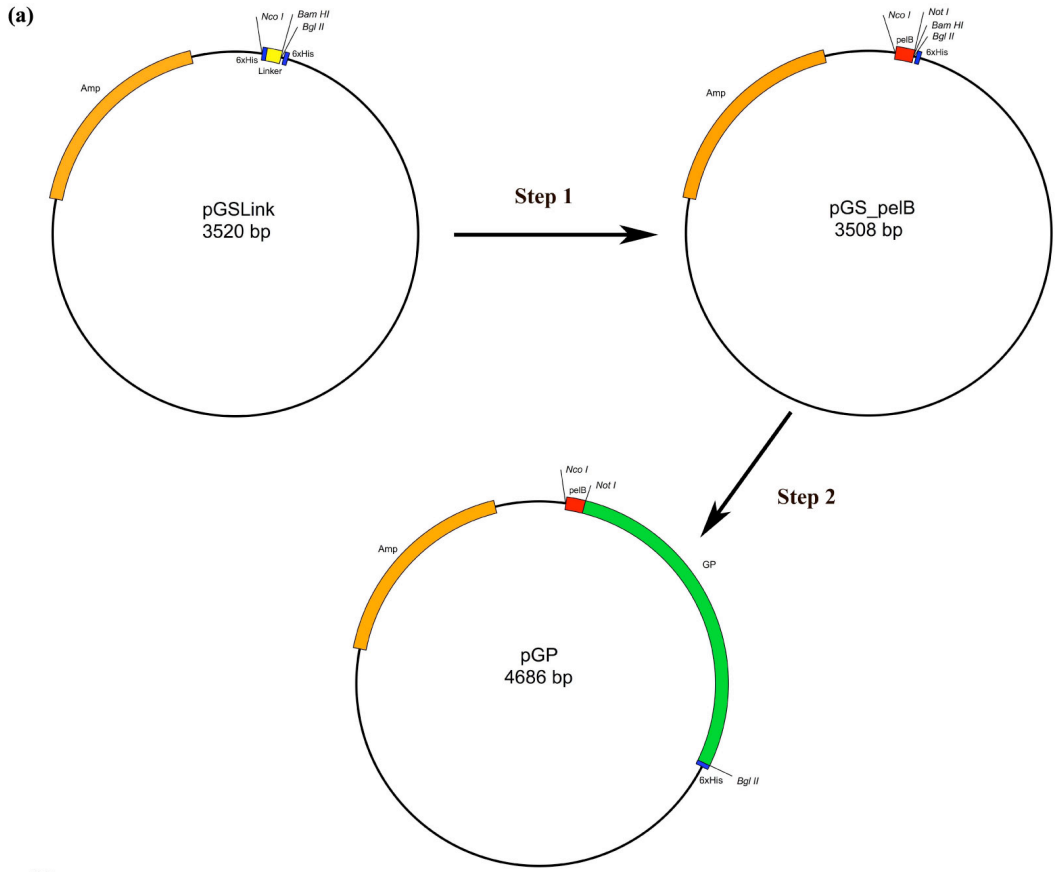
4.3.1 Ancestral gene synthesis and cloning

The MRCA of HRP A2, HRP C and SBP was inferred using *in silico* predictions and the phylogenetic tree of representatives of the plant peroxidases (Ryan, O'Connell, O'Fagain 2008). This ancestral sequence, referred to hereafter as the ancestral plant peroxidase (GP), was commercially synthesised by GENEART AG, Germany, see Figure 4.1. Optimised gene expression in prokaryotes was aided by codon choice/use.

The synthesised ancestral gene was delivered in a GENEART transport vector, pGA. A two step cloning strategy was employed to insert a pelB leader sequence and the GP gene sequence in frame in the expression vector, pGSLink; see Figure 4.2a for schematic.

Primers were designed for independent amplification of the pelB leader sequence from the pBR1 vector and the GP gene sequence from the pGA transport vector; see Tables 4.2 and 4.3 for plasmid descriptions and primer sequences respectively. Step 1: The pelB leader sequence was insesrted into the pGSLink expression vector via *Nco* I-*Not* I-*BamH* I cloning (pGS-pelB). Step 2: The GP gene sequence was then inserted into the resultant vector via *Not* I-*Bgl* II cloning (pGP). The resulting pGP plasmid had a C-terminal 6xHis tag downstream of the GP gene sequence for purification by metal affinity chromatography. The correct insertion of pelB-GP sequence in the modified pGSLink (pGP) was verified by sequencing (Eurofins/MWG Biotech, London) (Figure 4.2b), allowing for the optimisation of expression and purification of this ancestral plant peroxidase and preliminary biochemical characterisation of this novel enzyme.

(a)



(b)

```
  Nco I                                                                    Not I
CCATGGGCATGAAATACCTGCTGCCGACCGCTGCTGCTGGTCTGCTGCTCCTCGCTGCCCAGCCGGCGATGGCCGCGGCCGCCATGAAG
AATCTGTTTAACCTGTTTATGAAAATGATCATGAGCAGCATTAGCAAAGCGCTGTTCTTCTTTGTAATCCTGCTGGCCAGCTCTGCCAG
CGTAGACCTGCCAGGTGCGGGTGGTTATGGTGGCGACGACGATGACGATATCGCCTTTGCGGAGAACGAAGGTCATGGTCTGGCAATGA
ATTACTACTATAAGCTGCGTGATAGCTGTCCAGATAACGGCTCCCTGCTGAAGAATGTTAGCCCAGATTTTGATTCTCTGGAATCTGGC
GTCGCGTCTATTCTGACTCTGAATAATAACAAGAAACGCAATAGCGATAAGTACCTGCGCCAGCAACTGACGCCTGAGGCATGCGCGGA
AGAGATCATTCGTGAACAAGTGCAGCTGCTGTATAAACGTGATAAGAGCACCGCATTTAGCCTGCTGCGCCTGATCTTTCATGACTGCT
TTGTGGAGGGTTGTGATGCTTCTCTGCTGCTGGACTCCACCCGCCCGCGCATGCTGTCCGAGAAAGAGACTGATCGCCCGAATCTGTCT
TTTGGCCTGCGTGGTTTTCGCTATATCGATACGATTAAAGAAGCTCTGGAACGTGAACCTCAATGCCCGGGTGTATCTGTGTCTTGTGC
AGATATTCTGGTTCTGGCCGCGCGCGATGCAGTAGTGAGCCTGGGCGGCCCGTATATCCCGGTGAAAACCGGCCGCCGCGACGGTCGCA
AATCTCGCGCGGAAGACCTGGAACAGCATAACCTGCCAAGCCCGAACGAGAGCGTTAGCGTGGTTCTGGATCGCTTCGCATCCAAAGGC
CTGGACAATGTGCCGGACGTGGTTGCCCTGCTGGGTGCGGGCCACAGCGTTGGTAAGACGCATTGCAGCAAACTGGTCCATCGTTTTCG
CCTGTATAACTTCGTAGATACGGCTCTGAATCCGGACCCTCATGTAGAACATACGCTGCATAAGTGTCTGGACCCGCTGTGCCCGCCGT
CTGGCCAGAAAGGCCAGACCGATGCCGTACAGTATGTACGCAATGACCGCGGCACCCCGAATAAATTTGATAATAACTATTATCGCAAT
ATTCTGGACAACAAGGGCCTGCTGATTGTGGACCAGGAGCTGGCCACTGATAAACGCACGAAACCGTTTGTGAAGAAATTTGCAAAGAG
CTTCGCGAACGGCAACCAGGACTACTTCTTTAAAGAATTTGCACGCGCGATGACCGTGCTGGGTGAAATTAACGTACTGACTGGCACCA
AACGCCAGGGTGAGATCCGCAAACAATGTTATGTGGCCAATAAGCTGCATGTTGCGGTCCTGTCTGTTCTGGAAGAAGATGTAGAAGAA
GCCGTGGAGCTGCTGGCCGGTATGAGATCTCATCACCATCACCATCAC
                         Bgl II
```

Figure 4.2. Two step cloning. (a) Schematic representation of two step cloning strategy and (b) resultant sequence of pelB-GP insert. In red is the pelB leader sequence, green the ancestral plant peroxidase gene sequence (GP), blue the C-terminal 6xHis tag and restriction sites in black.

4.3.2 Expression and Purification of recombinant ancestral plant peroxidase

The theoretical molecular weight of this novel ancestral plant peroxidase was computed using the online Expasy Compute pI/MW Tool (http://www.expasy.ch/tools/pi_tool.html). The estimated value was approx. 54 kDa.

4.3.2.1 Selection of an E. coli expression strain for recombinant ancestral plant peroxidase

The pGP plasmid was transformed into three *E. coli* strains to ascertain the optimum host strain for expression of this novel enzyme; see Table 4.1. As a negative control, the three strains were also transformed with the pGSLink vector. Standard culturing conditions are described in Section 4.2.4.1. Cultures were either induced with IPTG (to a final concentration of 50 μM) or were not induced. Level of expression was determined based on the peroxidase activity (TMB absorbance) of the cell and supernatant fractions (Section 4.2.4.1); see Figure 4.3. No peroxidase activity was detected in the supernatant fractions. *E. coli* XL10Gold was found to be the optimum expression strain.

4.3.2.2 Optimistion of expression conditions for recombinant ancestral plant peroxidase

Culturing conditions for optimal expression of the GP in *E. coli* XL10Gold were then investigated. Temperature, incubation period and inducer concentration for expression of the ancestral plant peroxidse were assessed and determined. Previous studies on the expression of recombinant peroxidases revealed that repression of the production of the peroxidase was achieved by incubating the culture in the presence of 0.2 % (w/v) glucose until early exponential growth phase ($OD_{600\ nm} \approx 0.5$) was reached. This allowed for a burst in the production of the enzyme. This approach was applied here. Upon depletion, cultures were induced with varying concentrations of IPTG (0, 0.1, 0.5 and 1 mM) and grown at either 30 °C or 37 °C (Section 4.2.4.3). Samples were taken at 0, 2, 4

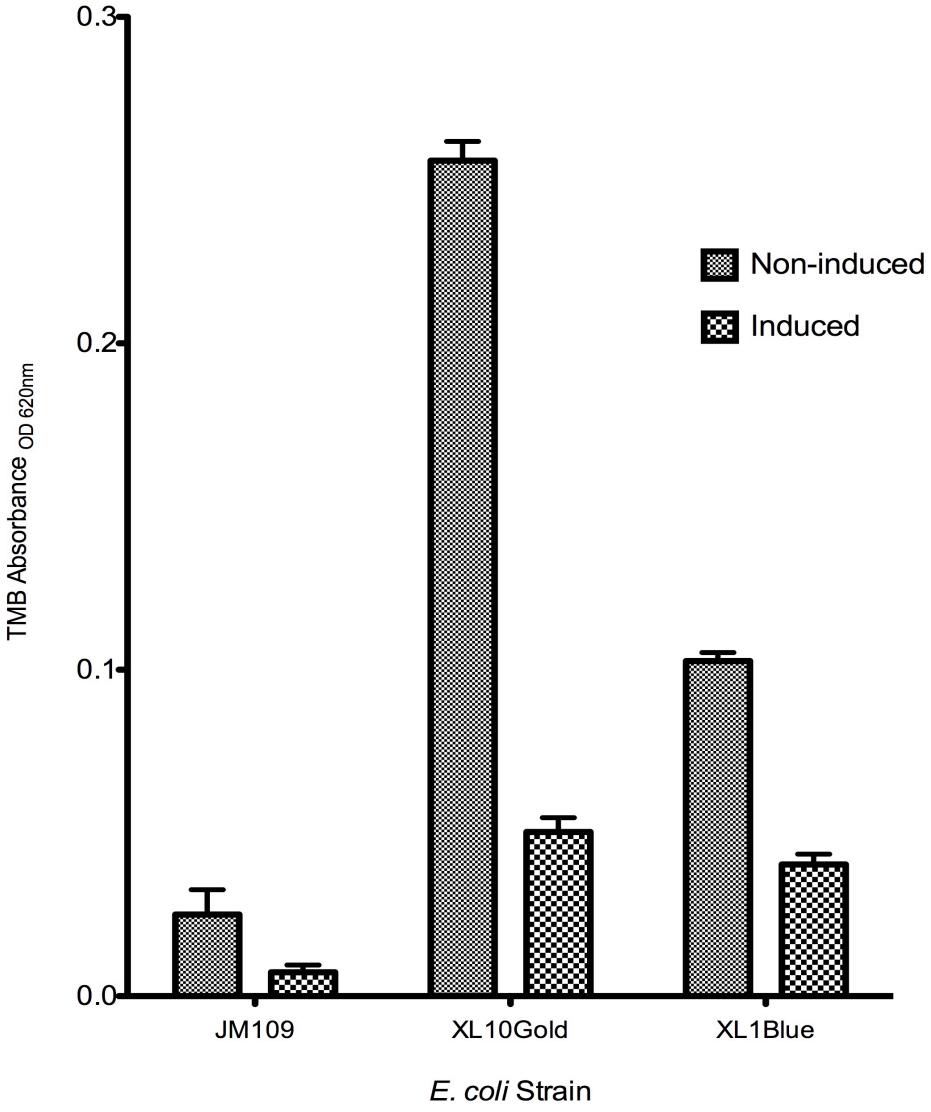


Figure 4.3. Optimal *E. coli* expression strain. *E. coli* strains expressing recombinant ancestral plant peroxidase were either induced with 50 μM IPTG or non-induced.

and 18 h time points. The amount of protein expressed in the soluble fraction of each sample was examined by SDS-PAGE (Section 3.2.4.3); see Figure 4.4. Expression of the ancestral plant peroxidase was not induced by IPTG. The optimal incubation temperature for expression was 30 °C with an incubation time of 18 h.

4.3.2.3 Purification of recombinant ancestral plant peroxidase by immobilised metal affinity chromatography (IMAC)

The recombinant ancestral peroxidase was expressed under optimum conditions (Section 4.2.4.4). Cell lysates were prepared as per Section 4.2.4.2 and subjected to Ni-NTA purification (Section 4.2.4.5). All fractions were analysed by SDS-PAGE (Figure 4.5). Two bands were evident in the elution fraction, one at the expected theoretical molecular weight for the GP, approx. 54 kDa, and one at approx. 100 kDa. This suggested the formation of a dimer; see Figure 4.5a. As such, the level of reducing agent (β-mercaptoethanol (β-ME)) in the SDS loading buffer (Appendix) was increased from 10 % (v/v) to 15 % (v/v) with the aim of denaturing the dimeric 100 kDa protein into its monomeric form. A single band at approx. 54 kD can be seen in Figure 4.5b. Purified fractions were pooled and concentrated for preliminary biochemical characterisation.

4.3.3 Characterisation of purified recombinant ancestral plant peroxidase

Protein concentration of the purified GP was determined by the standard BCA assay (Section 4.2.4.8). Optimised expression in *E. coli* XL10Gold yielded approx. 1.4 mg/L recombinant ancestral plant peroxidase. The Reinheitzahl (RZ) value (peroxidase purity number) was estimated at 0.4 (Section 4.2.4.9). Thermal and H_2O_2 tolerance of the GP compared to commercial HRP A2, HRP C and SBP were carried out, with a constant protein concentration of 0.1 mg/mL used in all investigations, where $n = 3$.

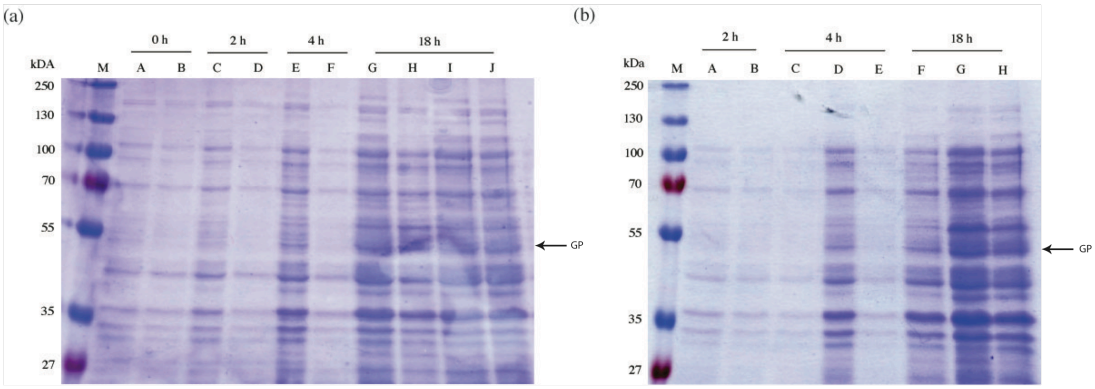


Figure 4.4. Recombinant ancestral plant peroxidase expression. SDS-PAGE analysis of the expression of the recombinant ancestral plant peroxidase (GP). (a) 37 °C; sample time points of 0, 2, 4 and 18 h are shown. Lane M: molecular weight standard, IPTG concentrations are shown in lanes A-J; A: 0 mM, B: 1 mM, C: 0 mM, D: 1 mM, E: 0 mM, F: 1 mM, G: 0.1 mM, H: 0 mM, I: 0.5 mM and J: 1 mM. (b) 30 °C; Sample time points of 2, 4 and 18 h are shown. Lane M: molecular weight standard, IPTG concentrations are shown in lanes A-H; A: 0 mM, B: 0.5 mM, C: 1 mM, D: 0 mM, E: 0.5 mM, F: 1 mM, G: 0 mM and H: 0.5 mM. Molecular weights of standards are shown in kDa. See additional Figure 2. (Appendix) for remaining sample time points and IPTG concentrations for both incubation temperatures.

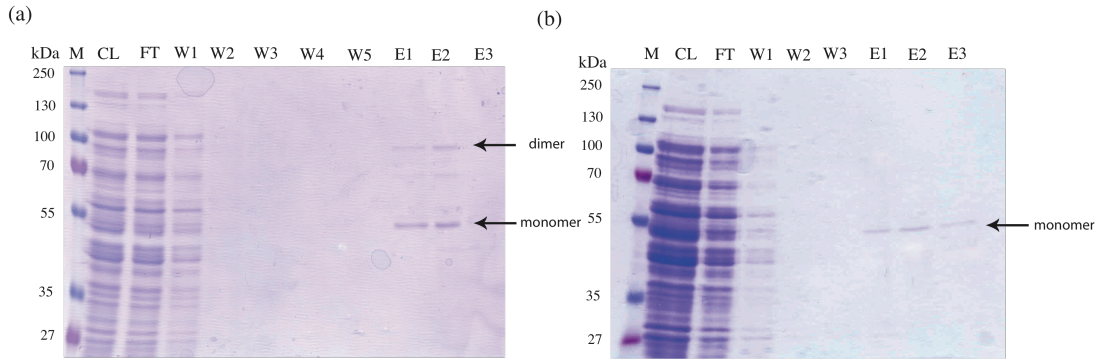


Figure 4.5. Purification of recombinant ancestral plant peroxidase. (a) Ni-NTA purification with 10 % (v/v)β-ME. Putative dimer (~100 kDa) and monomer (~ 54 kDa) present in elution fractions (E1-E2). (b) Ni-NTA purification with 15 % (v/v) β-ME. Monomer (~54 kDa) present in elution fractions (E1-E3). Lane M: molecular weight standards in kDa, CL: cell lysate, FT: flow through, W: wash fraction, E: elution fraction.

4.3.3.1 H_2O_2 stability of recombinant ancestral plant peroxidase

Although H_2O_2 is essential for peroxidation activity, in excess it can have a toxic effect. The H_2O_2 tolerance profile of each enzyme was determined as per Section 4.2.4.13. The recombinant ancestral plant peroxidase exhibits an increased stability to H_2O_2 with respect to its extant counterparts; see Figure 4.6. The C_{50} (H_2O_2 concentration at which 50 % residual enzyme activity is observed) of the GP, HRP A2, HRP C and SBP were approx. 40, 24, 17 and 7 mM respectively. An aliquot of H_2O_2-enzyme sample was taken to determine activity and, therefore, traces of H_2O_2 were not completely removed from the samples prior to assessing residual peroxidase activity. Carry-over amounts of of H_2O_2 could potentially affect the catalytic activity readings. However, under assay conditions, carry-over H_2O_2 is diluted further. The present method to determine oxidative stability was based on previous experimental methods within the group and from the literature (Hiner et al. 1996).

4.3.3.2 Thermal stability of recombinant ancestral plant peroxidase

4.3.3.2.1 Thermal profile

A thermal profile on all samples was performed to estimate the enzymes' approximate T_{50} (temperature at which 50 % residual enzyme activity is observed); see Figure 4.7. The recombinant ancestral plant peroxidase displayed poor to moderate thermal stability relative to the commercial peroxidases. Estimated T_{50} was 45 °C for the GP, whereas for commercial HRP A2, HRP C and SBP, T_{50} was 42 °C, 53 °C and 73 °C respectively.

4.3.3.2.2 Thermal inactivation

Thermal inactivation of the recombinant ancestral plant peroxidase at its T_{50} (45 °C) was assessed over period of 10 min. The percentage residual activity was fitted to a single exponential decay using the *Enzfitter* programme and the rate constant (*k*) was

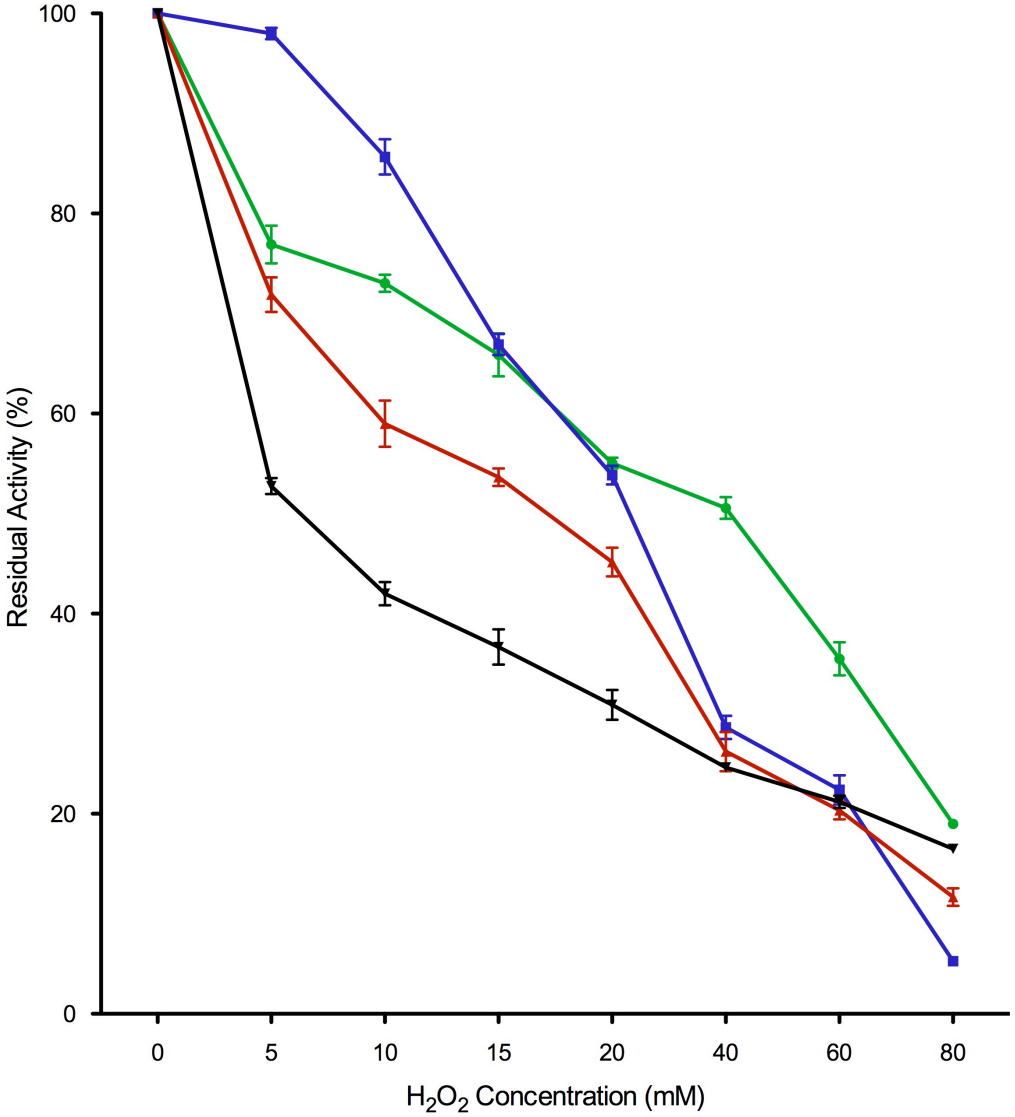


Figure 4.6. H_2O_2 tolerance profile. Profile of H_2O_2 stability of GP (green) compared to HRP A2 (blue), HRP C (red) and SBP (black), (n = 3).

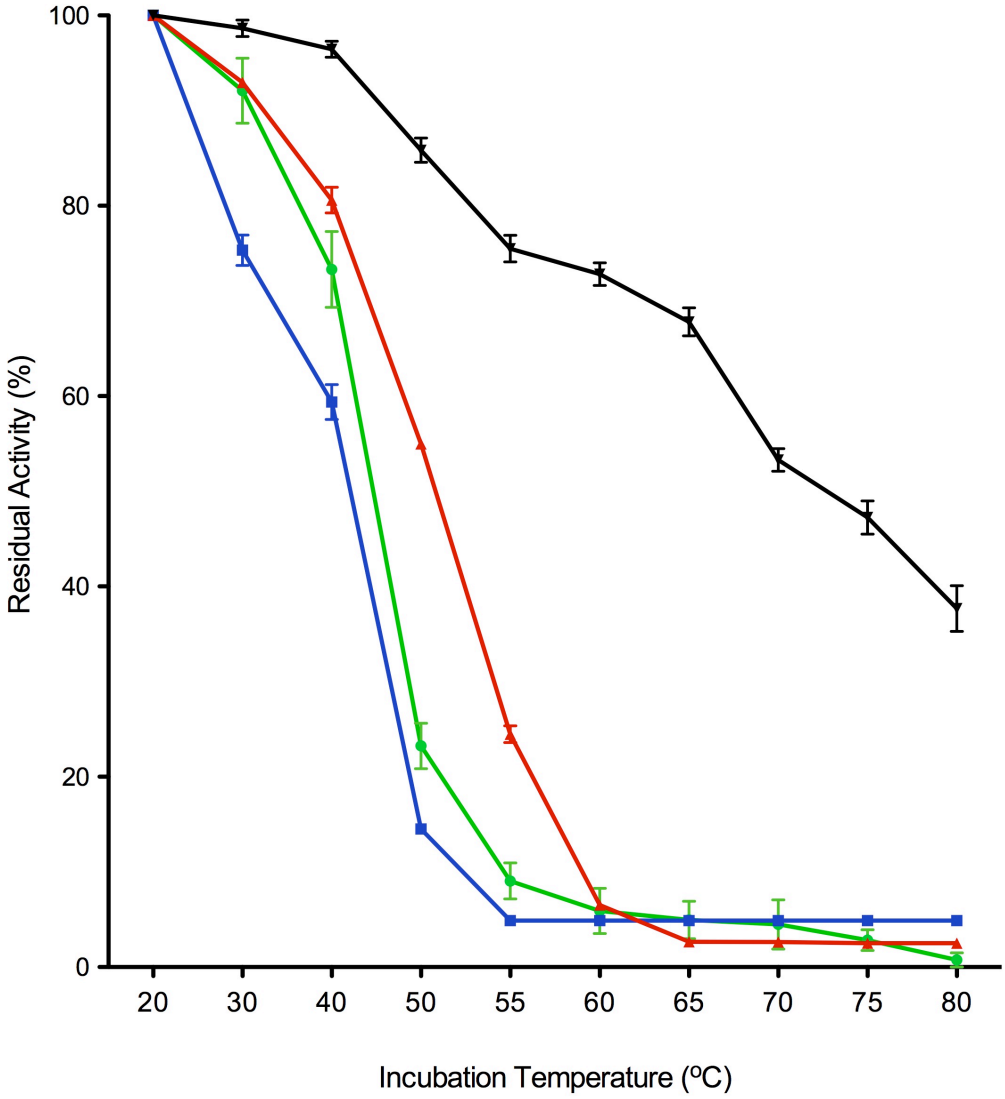


Figure 4.7. Thermal profile. Profile of thermal stability of GP (green) compared to HRP A2 (blue), HRP C (red) and SBP (black), (n = 3).

estimated, k-value = 0.0604 ± 0.0046 min^{-1}. The apparent half-life ($t_{1/2}$) was calculated, as per Section 4.2.4.14, to be 11.5 min. Thermal inactivation of commercial plant peroxidases was investigated at their respective T_{50} value. A summary of all thermal stabilities can be seen in Table 4.6 (n = 3).

4.3.3.3 ABTS kinetics of recombinant ancestral plant peroxidase

The extant plant peroxidases used in this study have previously been shown to follow Michaelis Menten kinetics, depicted by a hyperbolic curve of reaction rate versus substrate concentration (see Figure 4.8). Using the standard ABTS assay (as per Section 4.2.4.15), the kinetics of the ancestral plant peroxidase and HRP C were investigated. Non-Michaelis Menten kinetics were evident, with a sigmoidal curve of reaction rate versus substrate concentraion resulting (see Figure 4.8). Sigmoidal kinetics is often associated with multimeric enzymes, which may be the case with the recombinant ancestral plant peroxidase. A dimer was evident during purification anaylsis in Section 4.3.3. The GP enzyme may have been in a dimeric form during the preliminary kinetics experiments, which were performed in the absence of a reducing agent and/or substrate denaturant. Further detailed analysis into the kinetics and enzymatic characterisation of this ancestral plant peroxidase will provide greater insights into the evolution of this protein family.

Table 4.6: Thermal stability of plant peroxidases.

Enzyme	T_{50} (°C)	*k*-value (min^{-1})	SE (min^{-1})	$t_{1/2}$ (min)
GP	45	0.0604	± 0.0046	11.5
HRP A2	42	0.0611	± 0.0049	11.3
HRP C	53	0.0647	± 0.0051	10.7
SBP	73	0.0534	± 0.0077	13.2

Thermal inactivation of recombinant ancestral plant peroxidase (GP) and commercial plant peroxidases at their respective T_{50} (n = 3). The single exponential decay rate constant (k) and apparent half-life ($t_{1/2}$) for each enzyme were subsequently estimated with the aid of *Enzfitter* software (Biosoft, Cambridge, UK).

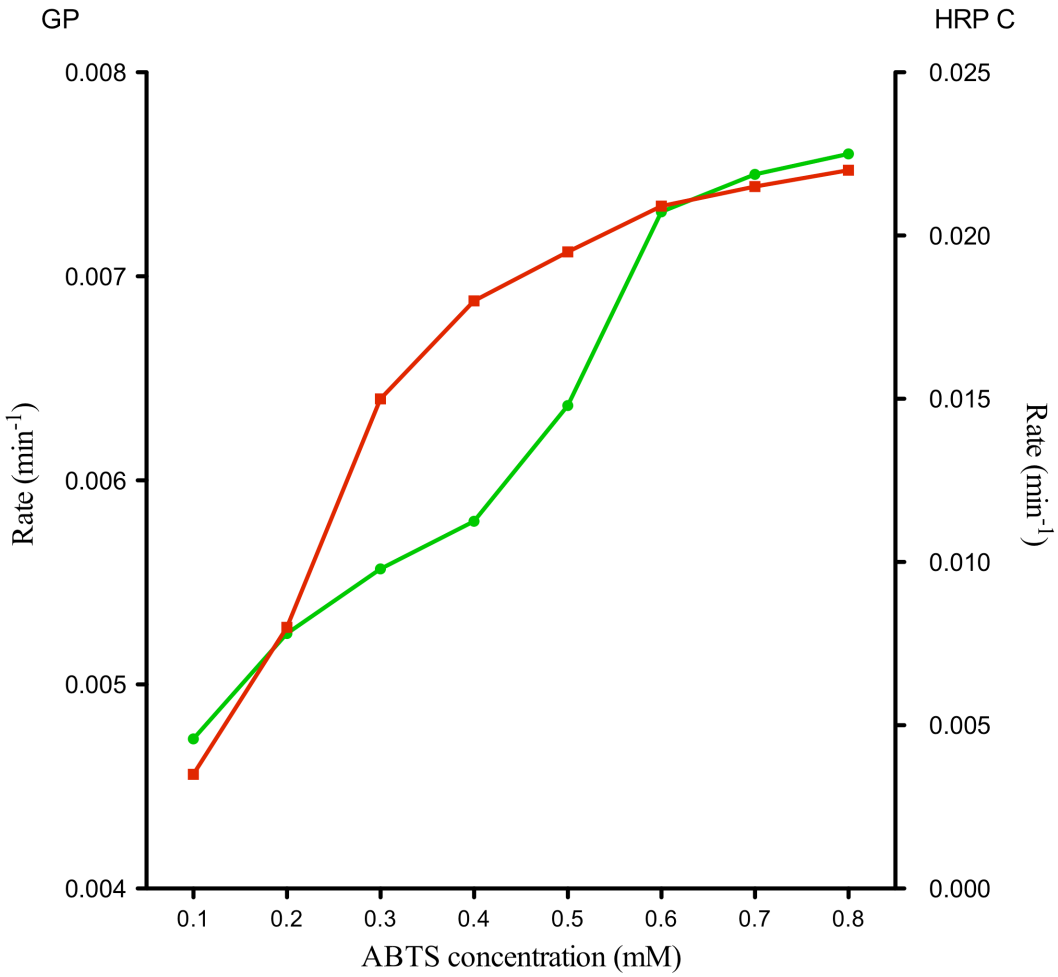


Figure 4.8. ABTS kinetics. Steady-state kinetics of recombinant ancestral plant peroxidsae (GP (green)) and extant HRP C (red). The rate ($\Delta A_{405\ nm}$/min) is the change in absorbance at 405 nm over the time of the assay (n = 3).

4.4 Discussion

Plant heme peroxidases are widely used in the biopharmaceutical and biotechnology sectors. Their operational potential is hampered by their thermal and oxidative stabilities. HRP A2, HRP C and SBP are members of the Class III secretory peroxidases, which are industrially important and all of which have their own set of desirable characteristics.

These three plant heme peroxidases, although closely related, have very diverse enzymatic stabilities – in terms of heat and H_2O_2 tolerance (McEldoon, Dordick 1996; Henriksen et al. 2001; Kamal, Behere 2003). With the recent advent of molecular biology and comparative genomics, knowing how these stably-diverse enzymes are related makes it possible to reconstruct and resurrect an hypothetical common ancestor of these Class III family members;this ancestral peroxidase can act as a tool to explore the evolution of the diverse enzymatic stabilities of this protein family.

The resurrection of such ancestral proteins may potentially recreate a protein that retains functional properties that may have been lost over millions of years by extant counterparts. From a fundamental science perspective, the study of ancestral proteins allows for a greater understanding of conditions millions of years ago, and gives a greater understanding about how proteins change over time. For example, studies on the evolutionary history of visual pigment proteins, incorporating the inference and synthesis of ancestral states of these proteins for the extinct archosaur, have revealed an ability to sense light under very low light levels, suggesting that ancient archosaurs, such as dinosaurs, existed under dim light conditions (Chang 2003).

In this study we have resurrected, in its active form, a 113 million year old plant heme peroxidase *de novo*. In comparison to its extant counterparts, HRP A2, HRP C and SBP, this ancient enzyme possesses moderate thermal stability and an increased tolerance to H_2O_2, which are desirable characteristics for industry. One might speculate that perhaps a more volitile environment existed in the Cretaceous period of the Mesozoic era and

that the characteristics exhibited by this 113 MYO protein may have been advantageous under such conditions.

The moderate thermal stability of this ancestral plant peroxidase is comparable to that of HRP A2. This may be attributable to the unstructured loops present only in the ancestral protein; see Figure 4.1b. These regions may be potential targets for future chemical modifications and/or protein engineering to increase the thermal stability of this ancient enzyme.

The operational stability of peroxidases in their many industrial and biomedical applications is limited by oxidative inactivation of the enzyme. These extant Class III peroxidases are readily inactivated in the presence of excess amounts of their primary substrate, H_2O_2. Much work has been carried out to decrease this oxidative degradation of peroxidases, including enzyme immobilisation (van de Velde, van Rantwijk, Sheldon 2001). Rational protein engineering to prevent/reduce this oxidative inactivation of peroxidases has been proposed (Valderrama, Ayala, Vazquez-Duhalt 2002). Increased tolerance to H_2O_2 will be beneficial in many sectors, predominantly in bioremediation/phenol clean-up. The ancestral plant peroxidase described here exhibits over a 1.5-fold increase in oxidative stability with respect to the most H_2O_2 tolerant enzyme in this study, HRP A2. This is a desirable operational characteristic over its extant peroxidase counterparts.

Preliminary characterisation of the enzyme kinetics of the ancestral plant peroxidase have revealed apparent sigmoidal (non-Michaelis Menten) kinetics atypical to the standard Michaelis Menten kinetics exhibited by its extant counterparts. Such sigmoidal kinetics is often associated with allosteric/multimeric enzymes that contain several interacting active sites; i.e. multi-subunit proteins with an active site in each subunit. Multimeric enzymes often give a sigmoidal rather than a hyperbolic (Michaelis Menten kinetics) curves due to the binding of one substate affecting the subsequent binding of additional substrates. This has been published previously for one notable example, the

multisubunit enzyme aspartate transcarbamoylase (Helmstaedt, Krappmann, Braus 2001).

This ancestral enzyme provides an excellent platform for future protein engineering and chemical modifications to investigate and enhance its catalytic and stability properties. Further biochemical characterisation of this 113 million year old plant peroxidase may unveil even more favourable traits for potential industrial and biomedical applications of this ancient enzyme.

Chapter 5

Discussion

5.1 Discussion

Comparative genomics is the anaysis and characterisation of both the differences and similarities between genomes. Comparative genomics and bioinformatics have made it possible to trace the evolutionary processes that are responsible for the divergence of genomes and ultimately protein functions. The results of comparative genomic analyses not only provide invaluable insights into how species have evolved but also allow for a greater understanding of the functions of genes and multigene families; for example, the evolution of protein specificity.

The exponential increase in the amount of gene and protein sequence data has placed molecular evolutionary biology at the forefront of biological sciences in the 21st century. Analysis of evolution at the molecular level considers the mutational processes that cause genetic variation in genes and genomes. These mutations can occur in several forms, including single nucleotide substitutions (e.g. synonymous/non-synonymous), insertion/deletion events and recombination. The fates of such mutations depend on the selective pressure acting on that region and also upon the effective population size. Genotypic variation may have an observable phenotypic effect, for example, a change in gene function/biological activity of the expressed protein.

Kimura proposed that the majority of evolutionary changes are due to the random fixation of neutral/nearly-neutral mutations, that is, through the process of random genetic drift (Kimura 1968; Kimura 1983). An alternative evolutionary mechanism is that of natural selection, where the overall outcome depends on the relative fitness of the genotypes competing for survival (reproductive success). Mutations that are deleterious to fitness are eventually eliminated from the population, known as purifying/negative selection. Desirable mutations that confer a selective/adaptive advantage are subject to the pressures of positive selection to become fixed in the continuing populations. Selectively neutral mutations, those that do not alter the overall fitness, are not influenced by the above pressures but drift through the population and may become fixed depending on the effective population size.

Following the process of gene duplication it has been observed that a number of alternative scenarios occur. The most important is the mechanism whereby one duplicate undergoes mutations that result in a new function for that copy of the gene , in a process known as 'neofunctionalisation'. While both duplicate copies are visible to natural selection in this scenario, the role of the original gene is maintained by one duplicate under a process of purifying selection. The second duplicate copy is free from the constraints of the original function and accumulates mutations at a higher rate as the selective pressure is relaxed on this copy. This duplicate copy can become subject to the pressures of positive selection/adaptive evolution, i.e. the aforementioned neofunctionalisation.

Patterns of gene duplication and neofunctionalisation are classically observed in large functionally diverse protein families. Using phylogenetics, the patterns of gene duplication can be traced. Molecular evolutionary analysis of this phylogenetic tree and the corresponding alignment allows for the identification of what the selective forces are acting upon gene duplicates. In this thesis, the MHP protein family has been used as a case study to determine the process by which functional specificity of enzymes/proteins occurs over millions of years.

In order to elucidate the possibility of using molecular evolution of a particular protein family to determine the substitutions responsible for shifts in specificity of proteins following gene duplication events, the MHP family were used as a case study. It was postulated that the detection of signatures of positive selection (adaptive evolution) across these enzymes would signify their observed functional shifts. The MHP have been classified into four main superfamilies; MPO, EPO, LPO and TPO. Previous studies have speculated on their relationship mainly from a functional viewpoint (Daiyasu, Toh 2000). This thesis (chapter 2), however, investigates the evolutionary processes at a molecular level that have influenced their functional diversity.

The phylogeny of mammals has been resolved and has made studies such as this one possible (Murphy et al. 2001). When estimating mammalian gene trees, one would

expect the branching pattern of the gene tree would be congruent with that of the resolved mammalian tree. However this is not always the case. The resolution of mammalian gene phylogenies encounters many challenges in determining the true evolutionary relationship. A number of systematic biases/pitfalls exist in phylogeny reconstruction that result in greater support for the incorrect evolutionary relationship. These include: inadequate phylogenetic signal resulting from mutationally saturated or highly conserved positions, poor modelling of the evolutionary processes and biases due to evolutionary rate variation among species. Unequal evolutionary rates across lineages may result in the LBA phenomenon, where lineages that have accumulated a greater number of mutations tend to inherit a basal position in the reconstructed phylogeny. This is the most prevalent systematic bias and a number of approaches have been proposed in order to overcome this phylogenetic artifact including: (i) increasing data sample size with the aim of breaking up long branch clusters, (ii) reconstructing the phylogeny on the slow evolving sites ('*Slow-Fast'* method), and (iii) using improved models of sequence evolution. Sampling more taxa to break up long branches is widely implemented to overcome this phenomenon. However a major shortcoming of this approach is that it is not suitable if all known taxa applicable to a study have been sampled when signatures of LBA still exist (Bergsten, Miller 2004). Another drawback of this approach is that addition of taxa may not be benefical and pose new problems, mainly due to their branch lengths and positioning on the tree (Poe, Swofford 1999). Inference methods such as maximum parsimony implement overly simplified models of sequence evolution and have been shown to be more susceptible to LBA (Philippe et al. 2005b). Inference methods incorporating improved models that take rate-heterogeneity into account are believed to be less sensitive (e.g. ML). However when poor model assumptions are made such inference methods may also suffer from this phylogenetics artifact (Bergsten 2005).

In chapter 2 of this thesis, the initial MHP phylogeny revealed the order of gene duplication events, where it is evident that there is an MPO-EPO-LPO MRCA arising from a gene duplication with the ancestor of extant TPO. This supports previous hypotheses from functional predictions. A further duplication event gave rise to the

common ancestor of MPO and EPO and the lineage leading to extant LPO. The final and most recent duplication was in the MPO-EPO MRCA leading to the extant MPO and EPO clades. This initial phylogeny had signatures of LBA within the functional clades, where the branching pattern of the species sampled was not in agreement with the resolved mammalian phylogeny. As such, a '*Slow-Fast*' approach was applied to overcome this artifact, whereby the most rapidly evolving sites were sequentially removed, and at each stage the representative phylogeny was estimated and its topology was compared with that of the expected resolved phylogeny. This approach resulted in the minimisation of LBA, yielding a highly supported fully resolved phylogeny for the MHP – estimated by two methods of phylogeny reconstruction, with both resultant phylogenies in agreement.

However, it is noteworthy that the '*Slow-Fast*' approach implemented in this thesis is not without its limitations. It is an approximate method for a complex evolutionary dynamic. Current methods for classifying the rate of evolution of sites in an alignment depend on the use of a *bona fide* phylogenetic tree. This means that data is classified on how well it fits this phylogeny. Such an approach is only suitable when there is high confidence/support that the phylogenetic relationship is true. For example, in mammalian systematics, the branching pattern of mammals is known, therefore, if the effects of LBA are minimised, the result is that branching pattern of the species sampled will be congruent with the known relationship of the species sampled. This is not the case for all datasets (e.g. bacterial datasets). If this initial tree is incorrect by removing those sites that conflict with the initial evolutionary relationship (e.g. most rapidly evolving sites), the resultant phylogeny is likely to be incorrect thus increasing the support for the wrong tree, resulting in somewhat of a circular argument. As such, this approach is not applicable when dealing with difficult phylogenies where there is no prior knowledge of the true evolutionary relationship of the species sampled. This circularity is something that should be addressed in the future with those developing methods for analysis of rates of substitution across sites.

Using this fully resolved phylogeny, positively selected sites have been identified, through the use of Bayesian estimations, unique to all four MHP: MPO, EPO, LPO and TPO. The activity of peroxidases is dependent on the presence and correct conformation of the heme group (Neves-Petersen et al. 2007). The conserved proximal histidines that are crucial for correct heme conformation are in close proximity to sites under positive selection in MPO, EPO and LPO. The majority of positively selected sites identified are in close proximity to catalytically important residues, suggesting that they may potentially be linked to functional shifts across the MHP. Many of these identified residues are also in neighbouring sites implicated in protein misfolding, loss of function or disease state.

Several claims have been made hypothesising the influence of positive selection on functional divergence but few have empirically illustrated the connection between evolutionary and functional shifts (e.g. Levassuer *et al.* 2006). A recent study has higlighted the need for experimental validation of *in silico* predictions of positively selected sites (Hughes 2008; Yokoyama et al. 2008). Yokoyama *et al.* (2008) investigated the evolution of phenotypic adaptations using visual pigments in vertebrates. They identified eight sites as being under positive selective pressure. Upon *in vitro* mutational analyses, these sites revealed no significant influence on the adaptation of rhodopsin sensitivity. Such a study emphasises the necessity to provide experimental evidence to support/validate computational analysis.

Chapter 2 of this thesis shows, from a molecular perspective, the relationships within the MHP multigene family and suggests that, following gene duplication, positive selection has led to the observed functional diversity among its family members. Using a cross-disciplinary approach in chapter 3, a small number of positively selected sites were analysed in detail in the MPO clade. This was done to investigate if there was a tangible connection between adaptive evolution and functional shift. Genetic abnormalities within the MPO gene are reflected by the disease state, MPO deficiency, which is marked by a loss/reduction in the enzyme's *in vivo* biological function (bactericidal immune response). Genotyping of a number of MPO deficient patients has revealed up

to five missense mutations (Nauseef, Brigham, Cogley 1994; Nauseef, Cogley, McCormick 1996; Romano et al. 1997; DeLeo et al. 1998; Ohashi et al. 2004; Persad et al. 2006; Goedken et al. 2007). Studies of the cellular fate of a number of these polymorphisms revealed aberrant processing of mature functional MPO (Nauseef, Cogley, McCormick 1996; Goedken et al. 2007). It is generally accepted that MPO's primary function is the generation of the potent cytotoxic agent HOCl. This differentiates MPO function from its MHP counterparts.

To test the hypothesis that positively selected amino acid residues are essential for the unique function of MPO a directed mutagenesis approach was applied. This was followed by biochemical analyses of the effects of mutating these residues at a molecular and phenotypic level, similar to the investigations of known causative mutations (Nauseef, Cogley, McCormick 1996; Goedken et al. 2007). Site directed mutagenesis was performed on two sites, Y500 and L504, identified from the *in silico* evolutionary study presented in chapter 2. These sites were chosen based on: (i) their closeness to the proximal heme ligand, His 502, and to the known MPO deficient mutants, R499C and G501S, and (ii) the confidence score from the *in silico* predictions.

From the results presented in chapter 3 it is evident that these positively selected residues within the MPO protein are involved in the observed protein functional shift. Pulse-chase and functional analyses revealed that there was a profound effect on the cellular fate and activity of MPO following mutation of these positively selected residues both independently and in combination (Y500F, L504T and double mutant Y500F-L504T). The synthesis of total MPO was not greatly influenced by the mutations - however the proteolytic processing into mature MPO was disrupted. To reiterate, the activity of peroxidases is dependent on the presence and correct conformation of the heme group (Neves-Peterson et al. 2007). There is an apparent maturational arrest in the biosynthetic processing of MPO between the two precursors, apoproMPO and proMPO. This is the point at which the prosthetic heme is normally incorporated. The impact of these mutations on the peroxidation and chlorination activity of MPO revealed the biological significance of the *in silico* predictions. MPO's peroxidation activity was

significantly reduced upon mutation. The unique property of MPO, production of the potent oxidant, HOCl, was obliterated following mutation of positively selected residues. This would suggest that these residues have been positively selected in the MPO lineage, as they accommodated the beneficial new function of chlorination activity and the production of HOCl.

This thesis has reviewed how molecular phylogenetics can be used to trace evolutionary processes such as gene duplications and adaptive evolution (neofunctionalisation) and how one can empirically test what role positive selection plays in functional divergence. Molecular phylogenetics, together with paleogenetics (paleomolecular biology), has the capacity to provide invaluable insights into the evolutionary path that produced observed functional differences. Paleomolecular biology is the resurrection and study of ancestral gene/protein states preduplication. Alternative to neofunctionalisation, the function of an ancestral state may be partitioned between two duplicates (i.e. subfunctionalisation). Subfunctionalisation is where both copies are subject to partial mutational degradation to maintain a complementry function that fulfils the essential function(s) of the ancestral gene. Such duplicate copies may be subject to subsequent neofunctionalisation (Mazet, Shimeld 2002). Ancestral functions may have also been lost over time due to nonsense mutations occurring within regulatory and/or protein coding regions. Ancestral protein resurrection can potentially connect ancient sequences to ancient molecular phenotypes, some of which may have desirable characteristics.

Enzymes are routinely used in biotechology applications (both biomedical and industrial). Protein engineering has applied both rational design and directed evolution to generate tailor-made proteins with the aim of increasing the operational efficacy of enzymes used in the biotechnology sector. Evolutionary biology provides crucial theoretical insights into protein function and can lend itself towards more efficient protein engineering approaches. Not only can paleomolecular biochemistry provide insights into mechanisms of evolution, it can potentially harness biological functions/processes that may have been lost over time, which may be beneficial in biomedical or industrial applications today.

The other major heme peroxidase family, the non-animal heme peroxidases (plant, bacteria and fungi) *specifically class III plant heme peroxidases,* were used as a case study in chapter 4 of this thesis, to determine if evolutionary theory (namely paleogenetics) can be used to design enzymes with desirable characteristics for industry. This family of enzymes has been extensively exploited in biopharmaceutical and biotechnology sectors through applications such as immunosensing technologies and industrial waste clean-up (Ryan, Carolan, O'Fagain 2006). The evolutionary relationship of this family of enzymes has been resolved (Duroux, Welinder 2003) and, as such, it is possible to reconstruct and resurrect an hypothetical common ancestor of these family members.

Chapter 4 presents the *de novo* resurrection of an active 113 million year old plant heme peroxidase. In comparison to its extant counterparts, which include the industrially relevant HRP and SBP enzymes, this ancient enzyme possesses moderate thermal stability and an increased tolerance of the extant enzymes' primary substrate, H_2O_2. The operational stability of its extant counterparts is limited by oxidative inactivation by H_2O_2 (substrate suicide inactivation). As such, these are desirable characterics for biomedical and industrial applications over the ancient protein's extant counterparts. Although only preliminary biochemical characterisations have been performed, it is evident that this ancient plant heme peroxidase provides an excellent test bed for future protein engineering and chemical modifications to enhance its catalytic and stability properties.

The heme peroxidases have been used as a case study for protein evolutionary dynamics in this thesis. The results presented in chapter 2 and 3 illustrate how the intersection of evolutionary theory with molecular biology techniques is a robust and innovative approach for allowing a greater understanding of the mechanisms of specificity and functional diversity in enzymes/proteins. These findings present a rare and unequivocal link between positive selection/adaptive evolution and protein functional shift. Chapter 4 highlights the potentials of evolutionary theory in designing tailor-made enzymes with

desirable characteristics for industry. The work presented here has provided an invaluable insight into the molecular evolution of proteins and the refinement of protein function and specificity.

Chapter 6

Bibliography

Abascal, F, R Zardoya, D Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104-2105.

Adey, NB, TO Tollefsbol, AB Sparks, MH Edgell, CA Hutchison. 1994. Molecular Resurrection of an Extinct Ancestral Promoter for Mouse L1. Proceedings of the National Academy of Sciences of the United States of America 91:1569-1573.

Aguinaldo, AM, JM Turbeville, LS Linford, MC Rivera, JR Garey, RA Raff, JA Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489-493.

Akaike, H. 1974. New Look at Statistical-Model Identification. Ieee Transactions on Automatic Control Ac19:716-723.

Ambrugger, P, I Stoeva, H Biebermann, T Torresani, C Leitner, A Gruters. 2001. Novel mutations of the thyroid peroxidase gene in patients with permanent congenital hypothyroidism. Eur J Endocrinol 145:19-24.

Andersson, E, L Hellman, U Gullberg, I Olsson. 1998. The role of the propeptide for processing and sorting of human myeloperoxidase. J Biol Chem 273:4747-4753.

Anisimova, M, JP Bielawski, Z Yang. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19:950-958.

Aratani, Y, H Koyama, S Nyui, K Suzuki, F Kura, N Maeda. 1999. Severe impairment in early host defense against Candida albicans in mice deficient in myeloperoxidase. Infect Immun 67:1828-1836.

Arnhold, J. 2004. Properties, functions, and secretion of human myeloperoxidase. Biochemistry (Mosc) 69:4-9.

Arnold, K, L Bordoli, J Kopp, T Schwede. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22:195-201.

Atkins, L and C.S. Bartsocas. 1974. Down's syndrome associated with two Robertsonian translocations, 45, XX, -15, -21, +t(15q21q) and 46, XX, -21, +t(21q21q). J Med Genet 11:306-309.

Azevedo, AM, VC Martins, DM Prazeres, V Vojinovic, JM Cabral, LP Fonseca. 2003. Horseradish peroxidase: a valuable tool in biotechnology. Biotechnol Annu Rev 9:199-247.

Bakalovic, N, F Passardi, V Ioannidis, C Cosio, C Penel, L Falquet, C Dunand. 2006. PeroxiBase: a class III plant peroxidase database. Phytochemistry 67:534-539.

Baldus, S, C Heeschen, T Meinertz, AM Zeiher, JP Eiserich, T Munzel, ML Simoons, CW Hamm. 2003. Myeloperoxidase serum levels predict risk in patients with acute coronary syndromes. Circulation 108:1440-1445.

Barman, TE. 1969. Enzyme Handbook vol. 1. New York: Springer-Verlag.

Benner, SA. 1988. Extracellular Communicator Rna. Febs Letters 233:225-228.

Benner, SA. 2002. The past as the key to the present: resurrection of ancient proteins from eosinophils. Proc Natl Acad Sci U S A 99:4760-4761.

Benner, SA, SO Sassi, EA Gaucher. 2007. Molecular paleoscience: systems biology from the past. Adv Enzymol Relat Areas Mol Biol 75:1-132, xi.

Benton, MJ, PC Donoghue. 2007. Paleontological evidence to date the tree of life. Mol Biol Evol 24:26-53.

Bergsten, J. 2005. A review of long-branch attraction. Cladistics 21:163-193.

Bergsten, J, BK Miller. 2004. Acilius phylogeny (Coleoptera : Dytiscidae), problems with long-branch attraction and morphological intersexual coevolution. Cladistics-the International Journal of the Willi Hennig Society 20:76-77.

Brennan, ML, MS Penn, F Van Lente, et al. 2003. Prognostic value of myeloperoxidase in patients with chest pain. N Engl J Med 349:1595-1604.

Brinkmann, H, H Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817-825.

Cai, W, JM Pei, NV Grishin. 2004. Reconstruction of ancestral protein sequences and its applications. Bmc Evolutionary Biology 4:-.

Cannarozzi, G, A Schneider, G Gonnet. 2007. A phylogenomic study of human, dog, and mouse. PLoS Comput Biol 3:e2.

Cao, Y, J Adachi, A Janke, S Paabo, M Hasegawa. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev Ecol Syst 28:437-446.

Carvalho, ASL, BS Ferreira, MT Neves-Petersen, SB Petersen, MR Aires-Barros, EP Melo. 2007. Thermal denaturation of HRPA2: pH-dependent conformational changes. Enzyme and Microbial Technology 40:696-703.

Cavalieri, EL, DE Stack, PD Devanesan, et al. 1997. Molecular origin of cancer: catechol estrogen-3,4-quinones as endogenous tumor initiators. Proc Natl Acad Sci U S A 94:10937-10942.

Chandrasekharan, UM, S Sanker, MJ Glynias, SS Karnik, A Husain. 1996. Angiotensin II-forming activity in a reconstructed ancestral chymase. Science 271:502-505.

Chang, BS, K Jonsson, MA Kazmi, MJ Donoghue, TP Sakmar. 2002. Recreating a functional ancestral archosaur visual pigment. Mol Biol Evol 19:1483-1489.

Chang, BSW. 2003. Ancestral gene reconstruction and synthesis of ancient rhodopsins in the laboratory. Integrative and Comparative Biology 43:500-507.

Chang, BSW, MJ Donoghue. 2000. Recreating ancestral proteins. Trends in Ecology & Evolution 15:109-114.

Chen, HJ, SW Row, CL Hong. 2002. Detection and quantification of 5-chlorocytosine in DNA by stable isotope dilution and gas chromatography/negative ion chemical ionization/mass spectrometry. Chem Res Toxicol 15:262-268.

Cheng, JD, RP Ryseck, RM Attar, D Dambach, R Bravo. 1998. Functional redundancy of the nuclear factor kappa B inhibitors I kappa B alpha and I kappa B beta. J Exp Med 188:1055-1062.

Chinen, A, Y Matsumoto, S Kawamura. 2005a. Reconstitution of ancestral green visual pigments of zebrafish and molecular mechanism of their spectral differentiation. Mol Biol Evol 22:1001-1010.

Chinen, A, Y Matsumoto, S Kawamura. 2005b. Spectral differentiation of blue opsins between phylogenetically close but ecologically distant goldfish and zebrafish. J Biol Chem 280:9460-9466.

Chung, WY, R Albert, I Albert, A Nekrutenko, KD Makova. 2006. Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. BMC Bioinformatics 7:46.

Clark, RA. 2000. Peroxidases: A historical overview of Milestones in Research on Myeloperoxidase. In: PE Petrides, WM Nauseef, editors. The peroxidase multigene family of enzymes: biochemical basis and clinical applications. Heidleberg: Springer-Verlag. p. 1-10.

Clark, RA, S Szot. 1981. The myeloperoxidase-hydrogen peroxide-halide system as effector of neutrophil-mediated tumor cell cytotoxicity. J Immunol 126:1295-1301.

Colas, C, PR Ortiz de Montellano. 2003. Autocatalytic radical reactions in physiological prosthetic heme modification. Chem Rev 103:2305-2332.

Conner, GE, M Salathe, R Forteza. 2002. Lactoperoxidase and hydrogen peroxide metabolism in the airway. Am J Respir Crit Care Med 166:S57-61.

Daiyasu, H, H Toh. 2000. Molecular evolution of the myeloperoxidase family. J Mol Evol 51:433-445.

Dale, DC, L Boxer, WC Liles. 2008. The phagocytes: neutrophils and monocytes. Blood 112:935-945.

Daugherty, A, JL Dunn, DL Rateri, JW Heinecke. 1994. Myeloperoxidase, a catalyst for lipoprotein oxidation, is expressed in human atherosclerotic lesions. J Clin Invest 94:437-444.

Davies, MJ, CL Hawkins, DI Pattison, MD Rees. 2008. Mammalian heme peroxidases: from molecular mechanisms to health implications. Antioxid Redox Signal 10:1199-1234.

Dayhoff, MO, RM Schwartz, BC Orcutt. 1978. A model of evolutionary change in proteins. Washington, DC: National Biomedical Research Foundation.

DeLeo, FR, M Goedken, SJ McCormick, WM Nauseef. 1998. A novel form of hereditary myeloperoxidase deficiency linked to endoplasmic reticulum/proteasome degradation. J Clin Invest 101:2900-2909.

deWit, JN, ACM vanHooydonk. 1996. Structure, functions and applications of lactoperoxidase in natural antimicrobial systems. Netherlands Milk and Dairy Journal 50:227-244.

dos Santos, WD, L Ferrarese Mde, A Finger, AC Teixeira, O Ferrarese-Filho. 2004. Lignification and related enzymes in Glycine max root growth-inhibition by ferulic acid. J Chem Ecol 30:1203-1212.

Dotsikas, Y, YL Loukas. 2004. Employment of 4-(1-imidazolyl)phenol as a luminol signal enhancer in a competitive-type chemiluminescence immunoassay and its

comparison with the conventional antigen-horseradish peroxidase conjugate-based assay. Analytica Chimica Acta 509:103-109.

Dowd, PF, LM Lagrimini. 1997. Examination of different tobacco (Nicotiana spp.) types under- and overproducing tobacco anionic peroxidase for their leaf resistance to Helicoverpa zea. Journal of Chemical Ecology 23:2357-2370.

Dunford, HB. 1999. Heme Peroxidases. New York: John Wiley and Sons Inc.

Duroux, L, KG Welinder. 2003. The peroxidase gene family in plants: a phylogenetic overview. J Mol Evol 57:397-407.

Edwards, AWF, Cavalli-Sforza. 1963. The reconstruction of evolution. Annals of Human Genetics 27.

Edwards, AWF, LL Cavvalli-Sforza. 1964. The reconstruction of evolution in Phenetic and Phylogenetic Classification,. London: Systematics Association.

Espelie, KE, VR Franceschi, PE Kolattukudy. 1986. Immunocytochemical localization and time course of appearance of an anionic peroxidase associated with suberization in wound-healing potato tuber tissue. Plant Physiol 81:487-492.

Felsenstein, J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Systematic Zoology 27:401-410.

Ferrand, M, V Le Fourn, JL Franc. 2003. Increasing diversity of human thyroperoxidase generated by alternative splicing. Characterized by molecular cloning of new transcripts with single- and multispliced mRNAs. J Biol Chem 278:3793-3800.

Fisher, RA. 1912. On an absolute criterion for fitting frequency curves. Messenger of Mathematics 41.

Force, A, M Lynch, FB Pickett, A Amores, YL Yan, J Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545.

Frey, A, B Meckelein, D Externest, MA Schmidt. 2000. A stable and highly sensitive 3,3 ',5,5 '-tetramethylbenzidine-based substrate reagent for enzyme-linked immunosorbent assays. Journal of Immunological Methods 233:47-56.

Fry, SC. 1986. Cross-Linking of Matrix Polymers in the Growing Cell-Walls of Angiosperms. Annual Review of Plant Physiology and Plant Molecular Biology 37:165-186.

Fugazzola, L, D Mannavola, MC Vigone, V Cirello, G Weber, P Beck-Peccoz, L Persani. 2005. Total iodide organification defect: clinical and molecular characterization of an Italian family. Thyroid 15:1085-1088.

Furtmüller, PG, M Zederbauer, W Jantschko, J Helm, M Bogner, C Jakopitsch, C Obinger. 2006. Active site structure and catalytic mechanisms of human peroxidases. Arch Biochem Biophys 445:199-213.

Gajhede, M, DJ Schuller, A Henriksen, AT Smith, TL Poulos. 1997. Crystal structure of horseradish peroxidase C at 2.15 A resolution. Nat Struct Biol 4:1032-1038.

Gaucher, EA, S Govindarajan, OK Ganesh. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451:704-707.

Gaucher, EA, JM Thomson, MF Burgan, SA Benner. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425:285-288.

Gerson, C, J Sabater, M Scuri, et al. 2000. The lactoperoxidase system functions in bacterial clearance of airways. Am J Respir Cell Mol Biol 22:665-671.

Goedken, M, S McCormick, KG Leidal, K Suzuki, Y Kameoka, JM Astern, M Huang, A Cherkasov, WM Nauseef. 2007. Impact of two novel mutations on the structure and function of human myeloperoxidase. J Biol Chem 282:27994-28003.

Goldschmidt, V, A Ciuffi, M Ortiz, D Brawand, M Munoz, H Kaessmann, A Telentil. 2008. Antiretroviral activity of ancestral TRIM5 alpha. Journal of Virology 82:2089-2096.

Gouy, M, M Chaussidon. 2008. Evolutionary biology: ancient bacteria liked it hot. Nature 451:635-636.

Grebski, E, C Peterson, TC Medici. 2001. Effect of physical and chemical methods of homogenization on inflammatory mediators in sputum of asthma patients. Chest 119:1521-1525.

Green, PS, AJ Mendez, JS Jacob, JR Crowley, W Growdon, BT Hyman, JW Heinecke. 2004. Neuronal expression of myeloperoxidase is increased in Alzheimer's disease. J Neurochem 90:724-733.

Gribaldo, S, H Philippe. 2002. Ancient phylogenetic relationships. Theor Popul Biol 61:391-408.

Grignaschi, VJ, AM Sperperato, D Catovsky, SA Farinati. 1963. [the Alkaline Phosphatase Reaction in the Neutrophil Leukocytes of the Human Blood.]. Prensa Med Argent 50:2567-2571.

Grisebach, H. 1981. Lignins, in secondary plant products. New York: Academic Press.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664-1674.

Gu, X, K Vander Velden. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18:500-501.

Guex, N, MC Peitsch. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714-2723.

Hampton, MB, AJ Kettle, CC Winterbourn. 1998. Inside the neutrophil phagosome: Oxidants, myeloperoxidase, and bacterial killing. Blood 92:3007-3017.

Hanson-Smith, V, B Kolaczkowski, JW Thornton. 2010. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. Mol Biol Evol.

Hansson, M, I Olsson, WM Nauseef. 2006. Biosynthesis, processing, and sorting of human myeloperoxidase. Arch Biochem Biophys 445:214-224.

Hartl, DL, AG Clark. 2007. Principles of population genetics, 4th Ed. U.S.: Sinauer Associates Inc.

Hazell, LJ, R Stocker. 1993. Oxidation of low-density lipoprotein with hypochlorite causes transformation of the lipoprotein into a high-uptake form for macrophages. Biochem J 290 (Pt 1):165-172.

Heinecke, JW. 1997. Pathways for oxidation of low density lipoprotein by myeloperoxidase: tyrosyl radical, reactive aldehydes, hypochlorous acid and molecular chlorine. Biofactors 6:145-155.

Heinecke, JW. 1999. Mechanisms of oxidative damage by myeloperoxidase in atherosclerosis and other inflammatory disorders. J Lab Clin Med 133:321-325.

Helmstaedt, K, S Krappmann, GH Braus. 2001. Allosteric regulation of catalytic activity: Escherichia coli aspartate transcarbamoylase versus yeast chorismate mutase. Microbiol Mol Biol Rev 65:404-421, table of contents.

Henriksen, A, O Mirza, C Indiani, K Teilum, G Smulevich, KG Welinder, M Gajhede. 2001. Structure of soybean seed coat peroxidase: a plant peroxidase with unusual stability and haem-apoprotein interactions. Protein Sci 10:108-115.

Hiner, AN, J Hernandez-Ruiz, MB Arnao, F Garcia-Canovas, M Acosta. 1996. A comparative study of the purity, enzyme activity, and inactivation by hydrogen peroxide of commercially available horseradish peroxidase isoenzymes A and C. Biotechnol Bioeng 50:655-662.

Hiner, AN, J Hernandez-Ruiz, JN Rodriguez-Lopez, MB Arnao, R Varon, F Garcia-Canovas, M Acosta. 2001. The inactivation of horseradish peroxidase isoenzyme A2 by hydrogen peroxide: an example of partial resistance due to the formation of a stable enzyme intermediate. J Biol Inorg Chem 6:504-516.

Holton, TA, D Pisani. 2010. Deep genomic-scale analyses of the metazoa reject coelomata: evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. Genome Biol Evol 2:310-324.

Huang, L, G Wojciechowski, PR Ortiz de Montellano. 2006. Role of heme-protein covalent bonds in mammalian peroxidases. Protection of the heme by a single engineered heme-protein link in horseradish peroxidase. J Biol Chem 281:18983-18988.

Huelsenbeck, JP, B Larget, RE Miller, F Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol 51:673-688.

Hughes, AL. 1999. Adaptive evolution of genes and genomes. New York: Oxford University Press.

Hughes, AL. 2008. The origin of adaptive phenotypes. Proc Natl Acad Sci U S A 105:13193-13194.

Hughes, T, DA Liberles. 2007. The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. J Mol Evol 65:574-588.

Ivics, Z, PB Hackett, RH Plasterk, Z Izsvak. 1997. Molecular reconstruction of Sleeping beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91:501-510.

Iwabata, H, K Watanabe, T Ohkuri, S Yokobori, A Yamagishi. 2005. Thermostability of ancestral mutants of Caldococcus noboribetus isocitrate dehydrogenase. FEMS Microbiol Lett 243:393-398.

Jermann, TM, JG Opitz, J Stackhouse, SA Benner. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature 374:57-59.

Johnson, NA. 2007. Darwinian Detectives: Revealing the Natural History of Genes and Genomes. USA: Oxford University Press.

Jones, DT, WR Taylor, JM Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-282.

Jong, EC, AA Mahmoud, SJ Klebanoff. 1981. Peroxidase-mediated toxicity to schistosomula of Schistosoma mansoni. J Immunol 126:468-471.

Josephy, PD. 1996. The role of peroxidase-catalyzed activation of aromatic amines in breast cancer. Mutagenesis 11:3-7.

Kamal, JK, DV Behere. 2003. Activity, stability and conformational flexibility of seed coat soybean peroxidase. J Inorg Biochem 94:236-242.

Karam, J, JA Nicell. 1997. Potential applications of enzymes in waste treatment. Journal of Chemical Technology and Biotechnology 69:141-153.

Keane, TM, CJ Creevey, MM Pentony, TJ Naughton, JO McLnerney. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol 6:29.

Keane, TM, TJ Naughton, JO McInerney. 2007. MultiPhyl: a high-throughput phylogenomics webserver using distributed computing. Nucleic Acids Res 35:W33-37.

Kimura, M. 1968. Evolutionary Rate at Molecular Level. Nature 217:624-&.

Kimura, M. 1983. The Neutral theory of molecular evolution. Cambridge: Cambridge Universoty Press.

Kimura, S, M Ikeda-Saito. 1988. Human myeloperoxidase and thyroid peroxidase, two enzymes with separate and distinct physiological functions, are evolutionarily related members of the same gene family. Proteins 3:113-120.

Klebanoff, SJ. 1970. Myeloperoxidase: contribution to the microbicidal activity of intact leukocytes. Science 169:1095-1097.

Klebanoff, SJ. 1991. Peroxidases in Chemistry and Biology. Boca Ratin, Florida: CRC PRESS.

Klebanoff, SJ. 1999. Myeloperoxidase. Proc Assoc Am Physicians 111:383-389.

Klebanoff, SJ. 2005. Myeloperoxidase: friend and foe. J Leukoc Biol 77:598-625.

Klebanoff, SJ, RW Coombs. 1996. Virucidal effect of stimulated eosinophils on human immunodeficiency virus type 1. AIDS Res Hum Retroviruses 12:25-29.

Knowles, DG, A McLysaght. 2009. Recent de novo origin of human protein-coding genes. Genome Res 19:1752-1759.

Koeffler, HP, J Ranyard, M Pertcheck. 1985. Myeloperoxidase: its structure and expression during myeloid differentiation. Blood 65:484-491.

Kohne, DE. 1970. Evolution of higher-organism DNA. Q Rev Biophys 3:327-375.

Kooter, IM, N Moguilevsky, A Bollen, LA van der Veen, C Otto, HL Dekker, R Wever. 1999. The sulfonium ion linkage in myeloperoxidase. Direct spectroscopic detection by isotopic labeling and effect of mutation. J Biol Chem 274:26794-26802.

Koua, D, L Cerutti, L Falquet, CJ Sigrist, G Theiler, N Hulo, C Dunand. 2009. PeroxiBase: a database with new tools for peroxidase family classification. Nucleic Acids Res 37:D261-266.

Lagrimini, LM. 1991. Wound-Induced Deposition of Polyphenols in Transgenic Plants Overexpressing Peroxidase. Plant Physiology 96:577-583.

Laird, CD, Mcconaug.Bl, BJ Mccarthy. 1969. Rate of Fixation of Nucleotide Substitutions in Evolution. Nature 224:149-&.

Lartillot, N, H Brinkmann, H Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7 Suppl 1:S4.

Lau, D, S Baldus. 2006. Myeloperoxidase and its contributory role in inflammatory vascular disease. Pharmacol Ther 111:16-26.

Lehrer, RI, MJ Cline. 1969. Leukocyte myeloperoxidase deficiency and disseminated candidiasis: the role of myeloperoxidase in resistance to Candida infection. J Clin Invest 48:1478-1488.

Lehrer, RI, J Hanifin, MJ Cline. 1969. Defective bactericidal activity in myeloperoxidase-deficient human neutrophils. Nature 223:78-79.

Levasseur, A, P Gouret, L Lesage-Meessen, M Asther, M Asther, E Record, P Pontarotti. 2006a. Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family. BMC Evol Biol 6:92.

Levasseur, A, P Gouret, L Lesage-Meessen, M Asther, E Record, P Pontarotti. 2006b. Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family. BMC Evol Biol 6:92.

Lewin, B. 2003. Genes VIII. Benjamin Cummings, USA.

Li, JY, F Gillard, A Moreau, JL Harousseau, C Laboisse, N Milpied, R Bataille, H Avet-Loiseau. 1999. Detection of translocation t(11;14)(q13;q32) in mantle cell lymphoma by flurescence *in situ* hybridization. Am J Pathol. 154(5):1449-1452.

Li, WH, DL Ellsworth, J Krushkal, BH Chang, D Hewett-Emmett. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. Mol Phylogenet Evol 5:182-187.

Li, WH, M Tanimura, PM Sharp. 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330-342.

Li, WH, CI Wu, CC Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150-174.

Long, M, E Betran, K Thornton, W Wang. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865-875.

Lunter, G. 2007. Dog as an outgroup to human and mouse. PLoS Comput Biol 3:e74.

Lynch, M, A Force. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

Mäder, M, R Füssl. 1982. Role of Peroxidase in Lignification of Tobacco Cells : II. Regulation by Phenolic Compounds. Plant Physiol 70:1132-1134.

Mader, SS. 1993. Biology. Dubuque: William C Brown Communications.

Malcolm, BA, KP Wilson, BW Matthews, JF Kirsch, AC Wilson. 1990. Ancestral Lysozymes Reconstructed, Neutrality Tested, and Thermostability Linked to Hydrocarbon Packing. Nature 345:86-89.

Marchetti, C, P Patriarca, GP Solero, FE Baralle, M Romano. 2004a. Genetic characterization of myeloperoxidase deficiency in Italy. Hum Mutat 23:496-505.

Marchetti, C, P Patriarca, GP Solero, FE Baralle, M Romano. 2004b. Genetic studies on myeloperoxidase deficiency in Italy. Jpn J Infect Dis 57:S10-12.

Martin, AC, AM Facchiano, AL Cuff, T Hernandez-Boussard, M Olivier, P Hainaut, JM Thornton. 2002. Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. Hum Mutat 19:149-164.

Matsuo, T, K Kuriyama, Y Miyazaki, et al. 2003. The percentage of myeloperoxidase-positive blast cells is a strong independent prognostic factor in acute myeloid leukemia, even in the patients with normal karyotype. Leukemia 17:1538-1543.

Mayr, E. 1942. Systematics and the origin of species. New York: Columbia University Press.

Mazet, F, SM Shimeld. 2002. Gene duplication and divergence in the early evolution of vertebrates. Curr Opin Genet Dev 12:393-396.

McEldoon, JP, JS Dordick. 1996. Unusual thermal stability of soybean peroxidase. Biotechnology Progress 12:555-558.

Meeusen, ENT, A Balic. 2000. Do eosinophils have a role in the killing of helminth parasites? Parasitology Today 16:95-101.

Messier, W, CB Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature 385:151-154.

Metcalfe, CL, M Ott, N Patel, K Singh, SC Mistry, HM Goff, EL Raven. 2004. Autocatalytic formation of green heme: evidence for H2O2-dependent formation of a covalent methionine-heme linkage in ascorbate peroxidase. J Am Chem Soc 126:16242-16248.

Mitra, SN, A Slungaard, SL Hazen. 2000. Role of eosinophil peroxidase in the origins of protein oxidation in asthma. Redox Rep 5:215-224.

Miyazaki, J, S Nakaya, T Suzuki, M Tamakoshi, T Oshima, A Yamagishi. 2001. Ancestral residues stabilizing 3-isopropylmalate dehydrogenase of an extreme

thermophile: experimental evidence supporting the thermophilic common ancestor hypothesis. J Biochem 129:777-782.

Moreira, D, H Philippe. 2000. Molecular phylogeny: pitfalls and progress. Int Microbiol 3:9-16.

Morgan, CC, NB Loughran, TA Walsh, AJ Harrison, MJ O'Connell. 2010. Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins. BMC Evol Biol 10:39.

Murphy, WJ, E Eizirik, WE Johnson, YP Zhang, OA Ryder, SJ O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614-618.

Mushegian, AR, JR Garey, J Martin, LX Liu. 1998. Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. Genome Res 8:590-598.

Nakagawa, T, T Ikemoto, T Takeuchi, K Tanaka, N Tanigawa, D Yamamoto, A Shimizu. 2001. Eosinophilic peroxidase deficiency: Identification of a point mutation (D648N) and prediction of structural changes. Hum Mutat 17:235-236.

Nascimento, AC, DR Guedes, CS Santos, M Knobel, IG Rubio, G Medeiros-Neto. 2003. Thyroperoxidase gene mutations in congenital goitrous hypothyroidism with total and partial iodide organification defect. Thyroid 13:1145-1151.

Nauseef, WM. 1986. Myeloperoxidase biosynthesis by a human promyelocytic leukemia cell line: insight into myeloperoxidase deficiency. Blood 67:865-872.

Nauseef, WM. 1989. Aberrant restriction endonuclease digests of DNA from subjects with hereditary myeloperoxidase deficiency. Blood 73:290-295.

Nauseef, WM, S Brigham, M Cogley. 1994. Hereditary myeloperoxidase deficiency due to a missense mutation of arginine 569 to tryptophan. J Biol Chem 269:1212-1216.

Nauseef, WM, M Cogley, S McCormick. 1996. Effect of the R569W missense mutation on the biosynthesis of myeloperoxidase. J Biol Chem 271:9546-9549.

Nauseef, WM, SJ McCormick, RA Clark. 1995. Calreticulin functions as a molecular chaperone in the biosynthesis of myeloperoxidase. J Biol Chem 270:4741-4747.

Nauseef, WM, I Olsson, K Arnljots. 1988. Biosynthesis and processing of myeloperoxidase--a marker for myeloid cell differentiation. Eur J Haematol 40:97-110.

Neves-Petersen, MT, S Klitgaard, AS Carvalho, SB Petersen, MR Aires de Barros, E Pinho e Melo. 2007. Photophysics and photochemistry of horseradish peroxidase A2 upon ultraviolet illumination. Biophys J 92:2016-2027.

Neyman, J. 1971. Molecular studies of evolution: A source of novel statistical problems in Statistical Decision Theory and Related Topics. New York: Academic Press.

Nicholls, SJ, SL Hazen. 2005. Myeloperoxidase and cardiovascular disease. Arterioscler Thromb Vasc Biol 25:1102-1111.

Nielsen, R, Z Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Nowak, MA, MC Boerlijst, J Cooke, JM Smith. 1997. Evolution of genetic redundancy. Nature 388:167-171.

O'Brien, PJ. 2000. Peroxidases. Chem Biol Interact 129:113-139.

O'Connell, MJ, JO McInerney. 2005. Gamma chain receptor interleukins: evidence for positive selection driving the evolution of cell-to-cell communicators in the mammalian immune system. J Mol Evol 61:608-619.

Ohashi, YY, Y Kameoka, AS Persad, F Koi, S Yamagoe, K Hashimoto, K Suzuki. 2004. Novel missense mutation found in a Japanese patient with myeloperoxidase deficiency. Gene 327:195-200.

Ohno, S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Ohta, T. 1993. An examination of the generation-time effect on molecular evolution. Proc Natl Acad Sci U S A 90:10676-10680.

Oliva, M, G Theiler, M Zamocky, D Koua, M Margis-Pinheiro, F Passardi, C Dunand. 2009. PeroxiBase: a powerful tool to collect and analyse peroxidase sequences from Viridiplantae. J Exp Bot 60:453-459.

Olsson, I, E Bulow, M Hansson. 2004. Biosynthesis and sorting of myeloperoxidase in hematopoietic cells. Jpn J Infect Dis 57:S13-14.

Orndorff, RC, N Stamm, S Craigg, et al. 2009. Divisions of geologic time - Major chronostratigraphic and geochronologic units. Stratigraphy 6:90-92.

Page, RD. 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14:819-820.

Page, RD, JA Cotton. 2002. Vertebrate phylogenomics: reconciled trees and gene duplications. Pac Symp Biocomput:536-547.

Page, RDM, EC Holmes. 1998. Molecular evolution, a phylogenetic approach. Oxford, UK: Blackwell Science Ltd.

Parra, A, ML Sanz, L Vila, I Prieto, I Dieguez, AK Oehling. 1999. Eosinophil soluble protein levels, eosinophil peroxidase and eosinophil cationic protein in asthmatic patients. J Investig Allergol Clin Immunol 9:27-34.

Passardi, F, N Bakalovic, FK Teixeira, M Margis-Pinheiro, C Penel, C Dunand. 2007a. Prokaryotic origins of the non-animal peroxidase superfamily and organelle-mediated transmission to eukaryotes. Genomics 89:567-579.

Passardi, F, G Theiler, M Zamocky, et al. 2007b. PeroxiBase: the peroxidase database. Phytochemistry 68:1605-1611.

Patterson, JW. 1985. Wastewater treatment technology, 2nd Ed. Ann Arbor, MI: Ann Arbor Science.

Pauling, L, E Zuckerkandl. 1963. Chemical Paleogenetics Molecular Restoration Studies of Extinct Forms of Life. Acta Chemica Scandinavica 17:9-&.

Persad, AS, Y Kameoka, S Kanda, Y Niho, K Suzuki. 2006. Arginine to cysteine mutation (R499C) found in a Japanese patient with complete myeloperoxidase deficiency. Gene Expr 13:67-71.

Petrides, P.E. and W.M. Nauseef. 2000. The peroxidase multigene family of enzymes: biochemical basis and clinical applications. Heidleberg: Springer-Verlag.

Philip, GK, CJ Creevey, JO McInerney. 2005. The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol 22:1175-1184.

Philippe, H, N Lartillot, H Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol 22:1246-1253.

Philippe, H, Y Zhou, H Brinkmann, N Rodrigue, F Delsuc. 2005a. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50.

Philippe, H, Y Zhou, H Brinkmann, N Rodrigue, F Delsuc. 2005b. Heterotachy and long-branch attraction in phylogenetics. Bmc Evolutionary Biology 5:-.

Poe, S, DL Swofford. 1999. Taxon sampling revisited. Nature 398:299-300.

Posada, D. 2003. Using MODELTEST and PAUP* to select a model of nucleotide substitution. Curr Protoc Bioinformatics Chapter 6:Unit 6 5.

Posada, D. 2006. ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. Nucleic Acids Res 34:W700-703.

Posada, D, KA Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Provine, WB. 2004. Ernst Mayr: Genetics and speciation. Genetics 167:1041-1046.

Puigbo, P, S Garcia-Vallve, JO McInerney. 2007. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics 23:1556-1558.

Rambaut, A. 1996. SE-AL Sequence alignment editor. Oxford.

Rao, R, JM Frederick, I Enander, RK Gregson, JA Warner, JO Warner. 1996. Airway function correlates with circulating eosinophil, but not mast cell, markers of inflammation in childhood asthma. Clin Exp Allergy 26:789-793.

Reiter, B. 1978. The lactoperoxidase-thiocyanate-hydrogen peroxide antibacterium system. Ciba Found Symp:285-294.

Reymond, A, G Meroni, A Fantozzi, et al. 2001. The tripartite motif family identifies cell compartments. EMBO J 20:2140-2151.

Reynolds, WF, E Chang, D Douer, ED Ball, V Kanda. 1997. An allelic association implicates myeloperoxidase in the etiology of acute promyelocytic leukemia. Blood 90:2730-2737.

Reynolds, WF, M Hiltunen, M Pirskanen, A Mannermaa, S Helisalmi, M Lehtovirta, I Alafuzoff, H Soininen. 2000. MPO and APOEepsilon4 polymorphisms interact to increase risk for AD in Finnish males. Neurology 55:1284-1290.

Reynolds, WF, J Rhees, D Maciejewski, T Paladino, H Sieburg, RA Maki, E Masliah. 1999. Myeloperoxidase polymorphism is associated with gender specific risk for Alzheimer's disease. Exp Neurol 155:31-41.

Rietsch, A, J Beckwith. 1998. The genetics of disulfide bond metabolism. Annu Rev Genet 32:163-184.

Rodrigues, C, P Jorge, JP Soares, I Santos, R Salomao, M Madeira, RV Osorio, R Santos. 2005. Mutation screening of the thyroid peroxidase gene in a cohort of 55 Portuguese patients with congenital hypothyroidism. Eur J Endocrinol 152:193-198.

Romano, M, P Dri, L Dadalt, P Patriarca, FE Baralle. 1997. Biochemical and molecular characterization of hereditary myeloperoxidase deficiency. Blood 90:4126-4134.

Ronquist, F, JP Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Roth, C, S Rastogi, L Arvestad, K Dittmar, S Light, D Ekman, DA Liberles. 2007. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. J Exp Zoolog B Mol Dev Evol 308:58-73.

Ruf, J, P Carayon. 2006. Structural and functional aspects of thyroid peroxidase. Arch Biochem Biophys 445:269-277.

Ryan, BJ, N Carolan, C O'Fagain. 2006. Horseradish and soybean peroxidases: comparable tools for alternative niches? Trends Biotechnol 24:355-363.

Ryan, BJ, MJ O'Connell, C O'Fagain. 2008. Consensus mutagenesis reveals that non-helical regions influence thermal stability of horseradish peroxidase. Biochimie 90:1389-1396.

Ryan, BJ, C O'Fagain. 2008. Effects of mutations in the helix G region of horseradish peroxidase. Biochimie 90:1414-1421.

Saitou, N, M Nei. 1986. The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. Japanese Journal of Genetics 61:611-611.

Saitou, N, M Nei. 1987. The Neighbor-Joining Method - a New Method for Reconstructing Phylogenetic Trees. Molecular Biology and Evolution 4:406-425.

Sakamaki, K, N Kanda, T Ueda, E Aikawa, S Nagata. 2000. The eosinophil peroxidase gene forms a cluster with the genes for myeloperoxidase and lactoperoxidase on human chromosome 17. Cytogenet Cell Genet 88:246-248.

Sakamaki, K, T Ueda, S Nagata. 2002. The evolutionary conservation of the mammalian peroxidase genes. Cytogenet Genome Res 98:93-95.

Sakharov, IY. 2001. Long-term chemiluminescent signal is produced in the course of luminol peroxidation catalyzed by peroxidase isolated from leaves of African oil palm tree. Biochemistry-Moscow 66:515-519.

Sakharov, IY, IS Alpeeva, EE Efremov. 2006. Use of soybean peroxidase in chemiluminescent enzyme-linked immunosorbent assay. Journal of Agricultural and Food Chemistry 54:1584-1587.

Sakharov, IY, AN Berlina, AV Zherdev, BB Dzantiev. 2010. Advantages of Soybean Peroxidase over Horseradish Peroxidase as the Enzyme Label in Chemiluminescent Enzyme-Linked Immunosorbent Assay of Sulfamethoxypyridazine. Journal of Agricultural and Food Chemistry 58:3284-3289.

Sanz, ML, A Parra, I Prieto, I Dieguez, AK Oehling. 1997. Serum eosinophil peroxidase (EPO) levels in asthmatic patients. Allergy 52:417-422.

Savenkova, MI, PR Ortiz de Montellano. 1998. Horseradish peroxidase: partial rescue of the His-42 --> Ala mutant by a concurrent Asn-70 --> Asp mutation. Arch Biochem Biophys 351:286-293.

Schmidt, HA, K Strimmer, M Vingron, A von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-504.

Schwarz, G. 1978. Estimating Dimension of a Model. Annals of Statistics 6:461-464.

Semple, CA, RS Devon, S Le Hellard, DJ Porteous. 2001. Identification of genes from a schizophrenia-linked translocation breakpoint region. Genomics 73(1):123-126.

Shi, YS, S Yokoyama. 2003. Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates. Proceedings of the National Academy of Sciences of the United States of America 100:8308-8313.

Shiu, SH, JK Byrnes, R Pan, P Zhang, WH Li. 2006. Role of positive selection in the retention of duplicate genes in mammalian genomes. Proc Natl Acad Sci U S A 103:2232-2236.

Sokal, RR, CD Michener. 1958. A Statistical Method for Evaluating Systematic Relationships. The University of Kansas Scientific Bulletin 38:1409-1438.

Stackhouse, J, SR Presnell, GM McGeehan, KP Nambiar, SA Benner. 1990. The ribonuclease from an extinct bovid ruminant. FEBS Lett 262:104-106.

Stendahl, O, BI Coble, C Dahlgren, J Hed, L Molin. 1984. Myeloperoxidase modulates the phagocytic activity of polymorphonuclear neutrophil leukocytes. Studies with cells from a myeloperoxidase-deficient patient. J Clin Invest 73:366-373.

Stewart, CB, JW Schilling, AC Wilson. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330:401-404.

Sugiyama, S, K Kugiyama, M Aikawa, S Nakamura, H Ogawa, P Libby. 2004. Hypochlorous acid, a macrophage product, induces endothelial apoptosis and tissue factor expression: involvement of myeloperoxidase-mediated oxidant in plaque erosion and thrombogenesis. Arterioscler Thromb Vasc Biol 24:1309-1314.

Sugiyama, S, Y Okada, GK Sukhova, R Virmani, JW Heinecke, P Libby. 2001. Macrophage myeloperoxidase regulation by granulocyte macrophage colony-stimulating factor in human atherosclerosis and implications in acute coronary syndromes. Am J Pathol 158:879-891.

Sun, HM, S Merugu, X Gu, YY Kang, DP Dickinson, P Callaerts, WH Li. 2002. Identification of essential amino acid changes in paired domain evolution using a novel combination of evolutionary analysis and in vitro and in vivo studies. Molecular Biology and Evolution 19:1490-1500.

Swofford, DL. 1993. Paup - a Computer-Program for Phylogenetic Inference Using Maximum Parsimony. Journal of General Physiology 102:A9-A9.

Tajima, T, J Tsubaki, K Fujieda. 2005. Two novel mutations in the thyroid peroxidase gene with goitrous hypothyroidism. Endocr J 52:643-645.

Tanaka, M, K Ishimori, M Mukai, T Kitagawa, I Morishima. 1997. Catalytic activities and structural properties of horseradish peroxidase distal His42 --> Glu or Gln mutant. Biochemistry 36:9889-9898.

Tauber, E, Y Herouy, M Goetz, R Urbanek, E Hagel, DY Koller. 1999. Assessment of serum myeloperoxidase in children with bronchial asthma. Allergy 54:177-182.

Tenovuo, JOaP, K. M. 1985. The Lactoperoxidase system: chemistry and biological significance. New York: Dekker.

Thomas, EL, KA Pera, KW Smith, AK Chwang. 1983. Inhibition of Streptococcus-Mutans by the Lactoperoxidase Anti-Microbial System. Infection and Immunity 39:767-778.

Thompson, JD, DG Higgins, TJ Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

Thomson, JM, EA Gaucher, MF Burgan, DW De Kee, T Li, JP Aris, SA Benner. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. Nat Genet 37:630-635.

Thornton, JW. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. Nature Reviews Genetics 5:366-375.

Thornton, JW, E Need, D Crews. 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. Science 301:1714-1717.

Ugalde, JA, BS Chang, MV Matz. 2004. Evolution of coral pigments recreated. Science 305:1433.

Undritz, E. 1966. [The Alius-Grignaschi anomaly: the hereditary constitutional peroxidase defect of the neutrophils and monocytes]. Blut 14:129-136.

Valderrama, B, M Ayala, R Vazquez-Duhalt. 2002. Suicide inactivation of peroxidases and the challenge of engineering more robust enzymes. Chem Biol 9:555-565.

van de Velde, F, F van Rantwijk, RA Sheldon. 2001. Improving the catalytic performance of peroxidases in organic synthesis. Trends Biotechnol 19:73-80.

Veitch, NC. 2004. Horseradish peroxidase: a modern view of a classic enzyme. Phytochemistry 65:249-259.

Wade, N. 1995. Method & madness; dead sure. The New York Times. New York.

Wagner, M, JA Nicell. 2002. Detoxification of phenolic solutions with horseradish peroxidase and hydrogen peroxide. Water Research 36:4041-4052.

Wang, J, A Slungaard. 2006. Role of eosinophil peroxidase in host defense and disease pathology. Arch Biochem Biophys 445:256-260.

Willson, SJ. 2005. Minimum evolution using ordinary least-squares is less robust than neighbor-joining. Bulletin of Mathematical Biology 67:261-279.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang, Z, S Kumar, M Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641-1650.

Yang, Z, R Nielsen, N Goldman, AM Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Yang, Z, WS Wong, R Nielsen. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107-1118.

Yap, YW, M Whiteman, NS Cheung. 2007. Chlorinative stress: an under appreciated mediator of neurodegeneration? Cell Signal 19:219-228.

Ye, XS, SQ Pan, J Kuc. 1990. Association of Pathogenesis-Related Proteins and Activities of Peroxidase, Beta-1,3-Glucanase and Chitinase with Systemic Induced Resistance to Blue Mold of Tobacco but Not to Systemic Tobacco Mosaic-Virus. Physiological and Molecular Plant Pathology 36:523-531.

Yokoyama, S, T Tada, H Zhang, L Britt. 2008. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proc Natl Acad Sci U S A 105:13480-13485.

Yoshida, K, P Kaothien, T Matsui, A Kawaoka, A Shinmyo. 2003. Molecular biology and application of plant peroxidase genes. Appl Microbiol Biotechnol 60:665-670.

Zamocky, M. 2004. Phylogenetic relationships in class I of the superfamily of bacterial, fungal, and plant peroxidases. Eur J Biochem 271:3297-3309.

Zederbauer, M, PG Furtmuller, S Brogioni, C Jakopitsch, G Smulevich, C Obinger. 2007. Heme to protein linkages in mammalian peroxidases: impact on spectroscopic, redox and catalytic properties. Nat Prod Rep 24:571-584.

Zhang, JZ, M Nei. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. Journal of Molecular Evolution 44:S139-S146.

Zhang, JZ, HF Rosenberg. 2002. Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. Proceedings of the National Academy of Sciences of the United States of America 99:5486-5491.

Zhu, G, GB Golding, AM Dean. 2005. The selective cause of an ancient adaptation. Science 307:1279-1282.

Publications

Manuscripts published:

- O'Connell, M.J., **Loughran N. B.**, Walsh, T. A., Donoghue M., Schmid K. and Spillane C. (2010) *"A phylogenetic approach to test for evidence of parental conflict or gene duplications associated with protein- encoding imprinted orthologous genes in placental mammals"*. Mammalian Genome. (In Press) . doi: 10.1007/s00335-010-9283-5. (Hard copy attached)

- Morgan, C. C., **Loughran, N. B.**, Walsh, T. A., and O'Connell, M. J. (2010) *"Positive selection neighbouring functionally essential sites and disease-implicated regions of mammalian reproductive proteins". BMC Evol. Biol.* **10**: 39. doi: 10.1186/1471-2148-10-39. (Hard copy attached)

- **Loughran, N. B.**, O'Connor B., Ó'Fágáin C., and O'Connell, M.J. (2008) *"The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions" BMC Evol. Biol.* **8**: 101. doi: 10.1186/1471-2148-8-101. (Hard copy attached)

Manuscripts submitted:

- **Loughran, N. B.**, Hinde, S. McCormick-Hill, S., Loughran, S. T., O'Connor, B., Ó Fágáin C., Nauseef, W. M., and O'Connell, M.J. (2010). *"Rational mutagenesis of human myeloperoxidase demonstrates the functional consequences of positively selected sites"*. In Review (MBE). (Electronic copy attached)

- McCole, R. B., **Loughran, N. B.**, Chahal, M., Fernandes, L. P., Fraternali, F., O'Connell, M. J., and Oakey, R. J. (2010) *"A case-by-case analysis of imprinted retrogene evolution"*. In Review (Evolution). (Electronic copy attached)

Mamm Genome
DOI 10.1007/s00335-010-9283-5

A phylogenetic approach to test for evidence of parental conflict or gene duplications associated with protein-encoding imprinted orthologous genes in placental mammals

Mary J. O'Connell · Noeleen B. Loughran · Thomas A. Walsh · Mark T. A. Donoghue · Karl J. Schmid · Charles Spillane

Received: 20 July 2010 / Accepted: 1 September 2010

Abstract There are multiple theories on the evolution of genomic imprinting. We investigated whether the molecular evolution of true orthologs of known imprinted genes provides support for theories based on gene duplication or parental conflicts (where mediated by amino-acid changes). Our analysis of 34 orthologous genes demonstrates that the vast majority of mammalian imprinted genes have not undergone any subsequent significant gene duplication within placental species, suggesting that selection pressures against gene duplication events could be operating for imprinted loci. As antagonistic co-evolution between imprinted genes can regulate offspring growth, proteins mediating this interaction could be subject to rapid evolution via positive selection. Supporting this, we detect evidence of site specific positive selection for the imprinted genes *OSBPL5* (and *GNASXL*), and detect lineage-specific positive selection for 14 imprinted genes where it is known that the gene is imprinted in a specific lineage, namely for: *PLAGL1, IGF2, SLC22A18, OSBPL5, DCN, DLK1, RASGRF1, IGF2R, IMPACT, GRB10, NAPIL4, UBE3A, GATM* and *GABRG3*. However, there is an overall lack of concordance between the known imprinting status of each gene (i.e. whether the gene is imprinted or biallelically expressed in a particular mammalian lineage) and positive selection. While only a small number of orthologs of imprinted loci display evidence of positive selection, we observe that the majority of orthologs of imprinted loci display high levels of micro-synteny conservation and have undergone very few *cis*- or *trans*-duplications in placental mammalian lineages.

Author contributions M.J.O'C, K.J.S and C.S designed research; M.J.O'C, N.B.L, T.A.W. and M.T.A.D. performed research; M.T.A.D, M.J.O'C, N.B.L, T.A.W., and CS prepared Results Figures and Tables, and M.J.O'C and C.S wrote the paper.

Noeleen B. Loughran, Mark T.A. Donoghue contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00335-010-9283-5) contains supplementary material, which is available to authorized users.

M. J. O'Connell · M. T. A. Donoghue · C. Spillane
Genetics and Biotechnology Lab, Department of Biochemistry and Biosciences Institute, University College Cork (UCC), Lee Maltings 2.10, Cork, Ireland
e-mail: mark.donoghue@nuigalway.ie

M. J. O'Connell · N. B. Loughran · T. A. Walsh
Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland
e-mail: noeleen.loughran@gmail.com

T. A. Walsh
e-mail: thomas.walsh26@mail.dcu.ie

M. J. O'Connell (✉) · N. B. Loughran · T. A. Walsh
Centre for Scientific Computing and Complex Systems Modelling (SCI-SYM), Dublin City University, Glasnevin, Dublin 9, Ireland
e-mail: mary.oconnell@dcu.ie

K. J. Schmid
Department of Plant Biology and Forest Genetics, Swedish University of Agricultural Sciences (SLU), P.O. Box 7080, 75007 Uppsala, Sweden
e-mail: karl@minzer-schmid.de

M. T. A. Donoghue · C. Spillane
Genetics and Biotechnology Lab, Centre for Chromosome Biology, National University of Ireland Galway, Galway, Ireland

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

Introduction

A wide range of evolutionary hypotheses have been proposed and debated to explain the selective pressures that drive the evolution of genomic imprinting (Haig and Trivers 1995; Hurst and McVean 1998), including hypotheses based on dosage compensation (Iwasa 1998), meiotic recombination (Pardo-Manuel de Villena et al. 2000), prevention of parthenogenesis (Varmuza and Mann 1994) and the widely cited kinship (or parental conflict) hypothesis (Moore and Haig 1991; Wilkins and Haig 2003).

Empirical evidence has been found in support of most of these hypotheses. For instance, the generation of viable parthenogenetic mice when the imprinted *Igf2/H19* and *Dlk1-Gtl2* regions are deleted provides experimental support for an imprinting barrier to asexual reproduction (Kono 2006; Kono et al. 2006; Kono et al. 2004). Imprinted regions also can display elevated frequencies of meiotic recombination suggesting a possible functional link between meiosis and imprinting (Lercher and Hurst 2003; Sandovici et al. 2006). Correlational studies have identified *cis*-linked features (e.g. such as repetitive elements and transposons) that are associated with imprinted loci, providing support for theories such as the host defense and meiotic recombination hotspot theories for the evolution of imprinted loci (Greally 2002; Luedi et al. 2007; Pask et al. 2009). Dosage compensation theories for the evolution of imprinting are based on associations between gene (or genome) duplication and gene dosage reduction via genomic imprinting (Iwasa 1998; Okamura et al. 2004; Walter and Paulsen 2003; Wood and Oakey 2006).

Kinship (or parental conflict) hypotheses for imprinting evolution propose that selection due to antagonistic co-evolution will drive mono-allelic expression of paternally derived alleles that increase maternal resource allocation to the offspring (growth enhancers), while growth inhibitors are predicted to be subject to mono-allelic expression from the maternally derived allele (Moore and Haig 1991). The parental conflict theory is supported by observations of expected overgrowth or undergrowth phenotypes when imprinted loci are disrupted in mammals (e.g. *GNASxl, Grb10, IGF2, Igf2R*) or plants (e.g. *MEDEA*) (Garnier et al. 2008; Haig 2004).

If kinship or parental conflicts involving antagonistic co-evolution between imprinted regulators of offspring growth are mediated via proteins encoded by imprinted genes, then such imprinted genes could be subject to rapid evolution via positive Darwinian selection (McVean and Hurst 1997). This possibility was previously investigated for the imprinted *Igf2* and *Igf2R* loci in the rat and mouse genomes where no evidence of antagonistic co-evolution was found (McVean and Hurst 1997). Since 1997, while some evidence has been found indicating positive Darwinian selection (or lineage accelerated evolution) for some imprinted genes in mammals (*Igf2R, KLF14*) (Parker-Katiraee et al. 2007; Smith and Hurst 1998), plants (*MEDEA*) (Spillane et al. 2007) and for non-imprinted homologs in placental fish (*IGF2*) (O'Neill et al. 2007). A recent investigation of the evolutionary history of mammalian imprinted genes found little evidence for positive selection, and concluded that most imprinted genes were under either purifying selection or relaxed constraints (Hutter et al. 2010).

In this study, using alignments of the true orthologs of 34 known protein-coding imprinted genes in mammals we analyzed their modes of evolution across both placental and non-placental mammalian lineages, and non-mammalian outgroups (e.g. chicken and fish). We have tested the imprinted orthologous gene sets for evidence of rapid-evolution of imprinted genes which could be indicative of antagonistic kinship or parental conflicts mediated by amino-acid changes. We also tested for any evidence of gene duplication events (within the clade where imprinting is known to have arisen) that could alter gene dosage within lineages of placental mammals and act as possible evolutionary drivers of imprinting regulation at such loci.

Materials and methods

Gene sequence data and sequence alignments

All of the gene data for mammalian imprinted genes were obtained from the imprinted gene catalog and from the ensembl database using the predefined orthologs. Data on the imprinted status of the known imprinted genes in human, mouse and other species were extracted from the Imprinted Gene Catalogue (http://igc.otago.ac.nz/). Coding DNA sequences for orthologs of these imprinted genes from placental mammal and the closest suitable outgroup species were extracted from the ensembl database (www.ensembl.org) using the predefined orthologs. These predefined orthologs are determined by performing a genome-wide reciprocal WUBlastp + SmithWaterman search of each imprinted gene across all completed genomes used in the analysis. Multiple sequence alignment (MSA) is then performed using the MUSCLE software (Edgar 2004) and the best reciprocal BLAST hits following the sequence similarity search. The longest alternative transcript in each

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

case was used. This approach was used to determine that these orthologs exist in single copy in each of these completed genomes, i.e. singletons. Only fully completed genomes were used in the analysis to ensure confidence in singletons identified.

Data on paralogy was gathered using the Ensembl Compara database (www.ensembl.org). The phylogenetic software implemented in the database construction is duplication aware (Vilella et al. 2009). The Compara database is a multispecies store of the results of genome-wide species comparisons and it houses information on ortholog and paralog predictions and protein family clusters for completed genomes (Altenhoff and Dessimoz 2009). These datasets were used for the analysis of gene duplication within the "imprinted clade" for each gene.

Outgroup species were chosen based on the phylogenetic distance to the placental mammals, and included gene sequences from chicken and fish. Out of a total of 63 known imprinted genes that code for proteins, 34 of these had greater than six species and included data from non-placental mammal species. The *GNASXL* gene provided one exception as, due to its rapid rate of evolution, the human and mouse orthologs cannot be aligned because they do not share enough sequence similarity. Therefore, in the case of *GNASXL* case only New- and Old-World Monkeys were analyzed, using the squirrel monkey as outgroup. Multiple sequence alignments (MSAs) for all data were created at the protein level using the default settings in ClustalW (Chenna et al. 2003), and gaps were subsequently entered into the nucleotide alignment where they were located in the protein-aligned data. All alignments were examined by eye for quality. In total it was possible to analyze 34 multiple sequence alignments of orthologs, allowing an analysis of 233 lineages in total (number calculated by summing the lineages present in each alignment). On average there were seven lineages per alignment.

Likelihood mapping

To determine whether the data had signal for phylogeny a Likelihood Mapping analysis was performed on each multiple sequence alignment. Each tree is disassembled into its constituent quartets and the support for each possible quartet is assessed. If the data contains phylogenetic signal then the likelihood of all three possible relationships for that quartet will be equally likely, these are represented by the three tips of the triangle, and the majority of the signal will be in these tip regions. Otherwise, the vertices and central region will be most heavily populated by supporting quartets.

Amino acid compositional bias

Amino acid composition bias was performed on all alignments using the TREE-PUZZLE package (Schmidt et al. 2002). TreePuzzle 5.2 performs a chi-square test that compares the amino acid composition of each sequence to the frequency distribution assumed in the General Time Reversible (GTR) and Jones Taylor Thornton (JTT) models (1992).

Micro-synteny analysis

Using online tools from the Ensembl website the genomic neighborhood of all imprinted genes was investigated in detail. The order and conservation of genes in imprinted regions were recorded for each of the 34 genes analysed.

Phylogeny reconstruction

Phylogeny reconstruction was carried out on the amino acid alignments throughout using MultiPhyl software (Keane et al. 2007). To determine whether the gene trees for each of the imprinted genes were concordant with the canonical mammalian species tree, phylogenetic trees for all multiple sequence alignments (MSAs) were created using the hierarchical Likelihood Ratio Test (hLRT) implemented in Multiphyl software(Keane et al. 2007). Maximum likelihood (ML) trees were inferred for each of the genes using the high-throughput phylogenomics web server, MultiPhyl (Keane et al. 2007). The Multiphyl software performs hierarchical likelihood ratio tests (hLRTs) of alternative models implemented. The ML tree for each individual dataset (34 in total) was reconstructed using the best-fit model of substitution in each case combined with the Nearest Neighbor Interchange (NNI) branch-swapping algorithm. All reconstructions were analyzed using 100 bootstrap replicates to determine support levels for all relationships. Following hLRT analysis the substitution model selected for each of the imprinted genes can be found in *Supplementary File 1*.

Shimodaira–Hasegawa statistical test of two phylogenetic trees

For each imprinted gene the closest upstream and downstream non-imprinted neighbours were identified. The phylogenies from these three genes and the species trees were compared using the Shimodaira–Hasegawa (SH) test implemented in the TreePuzzle 5.2 package (Schmidt et al. 2002). In each case the confidence set for each gene was examined and recorded. This analysis allowed us to determine whether there was a significant difference between the evolutionary histories of imprinted versus non-imprinted genes.

Gene birth and death analysis

In the three cases (*COPG2*, *GATM* and *IMPACT*) where the gene phylogenies and species phylogeny were not within the same confidence set, and other testable reasons for such incongruence had been tested, we finally tested if the conflict with the species tree was due to differential retention and loss of gene duplications. We applied the phylogeny reconciliation method in GeneTree (Page 1998). Using the pruned canonical species phylogeny as a template, all gene phylogenies were compared to this template. Using the software Genetree all gene duplications and losses required to derive the gene tree from the species tree were predicted. Reconciled trees demonstrate the gene birth and death events required in order to observe the pattern of retention of genes in extant genomes. To reduce the number of false positives predicted using this method we only used completed and well annotated genomes of high quality. This reduces the possibility that the losses we are deriving in the reconciled tree are due to undiscovered or unsequenced genes.

Analysis of selection pressures

Detection of positive Darwinian selection was conducted using a variety of models of codon sequence evolution using the PAML package v 3.15 (Yang 1997; Yang and Nielsen 1998). The Likelihood Ratio Test (LRT) was used to evaluate a variety of models of codon sequence evolution using the PAML package v3.15 (Yang 1997; Yang 1998; Yang et al. 2005; Zhang et al. 2005). The LRT proceeds by comparing nested models of sequence evolution. These models usually allow for variable D_n/D_s or w ratio among sites, along different branches of a phylogenetic tree or a combination of both. These models imply that there are a variety of classes of site in a given set of aligned sequences and the LRT provides a method of identifying the model that best describes the evolution of the set of sequences. In order to ascertain significance of difference in likelihood score we can compare the likelihood statistic $2\Delta lnL$ with X^2_v, where v is the number of degrees of freedom and corresponds with the number of free parameters. The models applied here have been described elsewhere and include the following: The neutral model of evolution M1, selection model M2, beta model M7, beta and omega model M8, beta and omega null model M8a (Yang 1997; Yang 1998; Yang et al. 2005; Zhang et al. 2005). We have applied strict criteria in our approach and only considered those alignments to have evidence of site specific positive selection if significant results were found ($p < 0.05$) for LRTs between: (1) M1 versus M2, (2) M7 versus M8, and (3) M8 versus M8a. Although this is a conservative approach we considered it necessary to avoid the detection/reporting of false positives. Lineage specific evolutionary rate variation was estimated using Model B and model M3 K = 2. In brief, M3 K = 2 allows for 2 site categories where the omega values are estimated from the data. Model B allows the user to specify a lineage as foreground and to calculate a Dn/Ds ratio for the foreground lineage separate from and distinct to the remainder of the phylogeny, this Dn/Ds ratio is allowed to vary above 1 and is estimated from the data. The LRT between the discrete model with 2 classes of site M3 K = 2 and Model B is performed with 2 degrees of freedom.

Results

Testing for differential selection pressures on imprinted and non-imprinted orthologs within the imprinting clade.

Many known imprinted genes in mammals can be traced to ancient gene duplication events, mostly predating the split of the fish and mammalian lineages (Hutter et al. 2010; Walter and Paulsen 2003). In the case of imprinted genes, following such ancient divergence of gene duplication events, one paralogous clade became imprinted in placental mammalian lineages while the other paralogous clade did not (Hutter et al. 2010). In this study we have focussed on the monophyletic clade (i.e. orthologs) that each imprinted locus resides within. We refer to this as the "imprinted clade" that relates to each imprinted locus (Fig. 1). This clade is of particular interest because not all orthologs within the clade are imprinted in all species

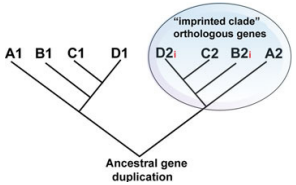


Fig. 1 Phylogenetic tree depicting the ancestral gene duplication events leading to orthologous "imprinted clade". The phylogeny depicts a gene duplication event in an ancestral non-imprinted gene, and subsequent speciation events into species *A*, *B*, *C* and *D*. The result is two paralogous groups containing on the *left* genes *A1–D1* and on the *right A2–D2*. In this figure two genes from the orthologous cluster on the right (*D2* and *B2*) are known to be imprinted. This "imprinted clade" of orthologous genes are the subject of this study. The imprinted clade of orthologous genes therefore consists of the monophyletic clade for which one or more of the mammalian species are imprinted

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

(i.e. some of the orthologs of imprinted genes have been shown to be biallelically expressed in some mammalian species) (Glaser et al. 2006). This study investigates the question of whether, within the imprinted clade, imprinted genes have been under differential selective pressures compared to their non-imprinted orthologous counterparts.

To ensure our datasets are robust, prior to analysis of the extent of heterogeneity in selective pressure across the "imprinted clade", we first performed a series of detailed analyses of our dataset of mammalian imprinted genes. In all cases we had greater than six species per gene. To determine the quality of signal in our "imprinted clade" datasets, on a gene-by-gene basis, we tested for (i) phylogenetic signal and data robustness, (ii) amino acid composition bias, (iii) evidence for orthology using synteny information and Compara database information at Ensembl (www.ensembl.org), and (iv) statistical robustness of the phylogenies.

The mammalian imprinted genes have robust phylogenetic signal

To test for alignment quality we used the likelihood mapping algorithm implemented in TREE-PUZZLE (Schmidt et al. 2002; Schmidt and von Haeseler 2007) whereby each phylogenetic tree is disassembled into its consituent quartets (four species), and the support for each possible quartet assessed. If the data contains phylogenetic signal then the likelihood of all three topologies (of four species) will be equally likely. Support for the three possible topologies for four species are represented by the three vertices of the triangle (Supplementary File 1). A significant proportion of quartets populating the edges and the central region of the plot would justify omitting the alignment from further analyses. This analysis of the phylogenetic signal in the datasets demonstrated that all of the 34 alignments have robust phylogenetic signal (Supplementary File 1).

Biased amino acid composition is not a major factor in the evolution of the "imprinted clade".

The amino acids in any set of homologous proteins can differ due to the nucleotides available at a particular position (G + C content). Such amino acid composition bias is directly correlated with nucleotide bias and can have the effect of erroneously indicating relatedness (Foster et al. 1997). The analysis of amino acid composition bias in the dataset indicates that only four lineages showed amino acid composition bias. These were from four different genes (*ASB4*: Opossum, *CD81*: Dog, *OSBPL5*: Chicken, and *PLAGL1*: Fish—see Supplementary File 2). Therefore, it can be concluded that biased amino acid composition is not a major contributing factor to the differential evolution of the vast majority of genes within the "imprinted clade" (Chang and Campbell 2000; Foster et al. 1997; Inagaki and Roger 2006).

Mammalian imprinted genes in the dataset have not undergone ancestral *trans*- or *cis*- duplications within the "imprinted clade" (i.e. they are true orthologs).

Gene duplication has been proposed as a possible evolutionary driver of imprinting in mammalian genomes (Walter and Paulsen 2003), and recent studies indicate that imprinted genes diverged early during vertebrate evolution (e.g. in many instances prior to divergence of the chicken lineage from mammals), with duplications leading to outparalogs that remain conserved in the genomes of extant mammals (Hutter et al. 2010). We hypothesised that imprinted genes could have undergone further ancestral duplication/loss events leading to differential retention of duplicates in separate mammalian lineages within the "imprinted clade".

To determine whether mammalian genomes contained any duplicates (in *trans* or *cis* chromosomal contexts) of the "imprinted clade" genes, we searched for evidence of retained gene duplicates in the genomes of all the species in our dataset using the Compara phylogeny database at Ensembl (www.ensembl.org). In total we find only four genes in the dataset where there is evidence of gene duplication and retention of that duplicate in mammalian species; i.e. *MEST*, *INS*, *DIO3* and *IMPACT*, which indicates that the remaining 88% of the 34 imprinted loci analysed are single gene orthologs. The complete set of results from this analysis are provided in *Supplementary File 3*.

In addition, to determine whether any imprinted genes could have undergone duplications followed by differential (yet balanced) retention and loss in different mammalian lineages (leading to singleton status), we employed the Ensembl database (www.ensembl.org). We examined the genomic neighbourhoods of all genes and species in the dataset. We observed complete conservation of microsynteny for imprinted gene loci across all eutherian mammals. This is highly suggestive of limited rearrangements of these genomic regions within mammalian lineages (see *Supplementary File 4* for summary of the synteny analysis). However, we do observe that (i) the clusters can be reverse in some eutherian species (but the order within the cluster is retained), and (ii), in non-eutherian lineages the genomic neighbourhood can be completely different. Together this pattern may suggest that before eutherian mammals these genes were more dispersed throughout the genome (e.g. *GATM* and *RASGRF1* in *Supplementary File 4*), and that after eutherian divergence these genes have been under stronger pressure to retain their clustering. Taken together, these results indicate that following the emergence of imprinting, subsequent gene duplication and retention within the "imprinted clade" is rare.

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

Phylogenetic gene trees for imprinted genes and flanking genes are congruent with the species tree in all but three cases.

To test for evidence of possible *cis*-duplications, gene trees were generated for every imprinted gene in the dataset. In all cases the gene trees contained some minor anomalies in comparison to the species phylogeny (see Fig. 2 for the phylogeny of species used in this analysis). To determine if anomalous patterns in each gene phylogeny were significantly different to the known species phylogeny, we employed the SH test. If an imprinted locus had undergone local *cis*-duplication followed by loss of the progenitor gene and retention of the duplicate, the divergence of the duplicate copy could account for any significant difference between the gene and species phylogenies. Hence, for every imprinted gene in the dataset, we generated gene phylogenies for the genes upstream and downstream of each imprinted locus. To determine if the imprinted genes had evolutionary histories (e.g. due to *cis* duplication and loss events) significantly different to their neighbouring genes, using the SH test, we compared all resultant gene phylogenies to the canonical species phylogeny. The results of the SH statistical test demonstrate that the species phylogeny was within the confidence set of phylogenies for the vast majority of imprinted genes and the neighbouring genes. See *Supplementary File 5* for a summary of the SH test results.

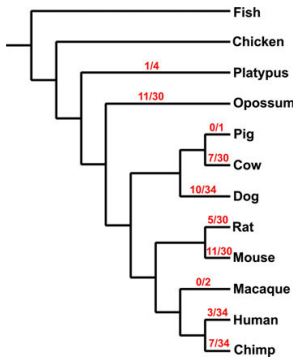


Fig. 2 Summary of positive selection detected in various lineages for imprinted clade orthologs. For all 34 genes in the imprintome dataset we have depicted on the canonical species phylogeny the results of positive selection per lineage. The *numbers in red* on each branch depict the number of genes with evidence of positive selection for that species as a function of the number of gene sets available for the species

Overall, these results propose that despite gene duplication and losses being evident prior to the emergence of imprinted genes, following the origin of imprinting, gene duplication and loss events are relatively rare in the imprinted clades.

Our SH analyses have identified three possible exceptions where differential gene duplication and loss may have occurred since imprinting emerged. The three genes are: *COPG2*, *GATM*, and *IMPACT* (*Supplementary File 5*). The phylogenetic trees for these three genes produce significantly different topologies to the neighbouring non-imprinting genes and to the canonical species phylogeny (*Supplementary File 5*). We have reconciled the gene and species phylogenies for these three genes where our results are suggestive of extensive gene duplications and losses (*Supplementary File 6*). However, the high numbers of duplications and losses necessary to reconcile the gene and species trees for *COPG2*, *GATM* and *IMPACT* could be an artefact of the analysis.

However, there is a significant difference between the gene tree and species tree when we apply the SH statistic in these three cases. Hence, we have used the gene tree for the selection analysis. As all three genes exist in single copy, a more likely scenario for the significant difference between the gene and species trees for these three genes may be a lineage sorting effect rather than an unreasonably large number of gene duplications and losses.

Overall, the results demonstrate that the mammalian imprinted genes are more likely to remain as single gene orthologs within mammalian lineages, and they retain strong phylogenetic signal that is not significantly different to the canonical species phylogeny.

OSBPL5 and *GNASXL* display evidence of site-specific positive Darwinian selection

To determine whether any of the 34 mammalian imprinted loci could be rapidly evolving due to positive selection, all orthologous mammalian imprinted genes were analysed using Maximum Likelihood (ML) site-specific models of evolution. Likelihood Ratio Tests (LRTs) were performed between pairs of models that allow for site-specific evolution, namely M1 versus M2, M7 versus M8, and M8 versus M8a (Swanson et al. 2003; Yang 1997; Yang 1998; Yang et al. 2005; Zhang et al. 2005). For the 34 mammalian imprinted genes tested using site-specific models of evolution, only two imprinted genes (*OSBPL5* and *GNASXL*) displayed a significant level of site-specific positive selection (Table 1). The two sites under positive selection in the oxysterol binding protein-like 5 (*OSBPl5*) are located at amino-acid positions 194(S) and 869(H). We investigated the position of these residues for possible

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

Table 1 Analysis of the significance of fit of site-specific models of evolution

Gene	Outgroups	Codeml M1 versus M2, M7 versus M8 and M8a versus M8
TP73	Fish	−
COMMD1	Fish	−
PLAGL1	Fish	−
SLC22A2	Opossum	−
GRB10	Fish	−
CALCR	Chicken	−
SGCE	Fish	−
PON2	Fish	−
ASB4	Fish	−
DLX5	Chicken	−
CPA4	Chicken	−
MEST	Fish	−
COPG2	Fish	−
CTNNA3	Chicken	−
IGF2	Fish	−
INS	Fish	−
PHEMX	Chicken	−
CD81	Fish	−
SLC22A18	Fish	−
NAP1L4	Fish	−
OSBPL5	Fish	+
WT1	Chicken	−
DCN	Chicken	−
HTR2A	Chicken	−
DLK1	Fish	−
DIO3	Chicken	−
UBE3A	Chicken	−
GABRB3	Fish	−
GABRG3	Fish	−
GATM	Fish	−
RASGRF1	Fish	−
IMPACT	Fish	−
SLC38a4	Fish	−
IGF2R	Fish	−
GNASXL	Squirrel monkey	+
Total	34	2

The *left* column represents the genes in the dataset. The *central* column represents those non-mammalian species for which a homolog of the gene exists. The exception is *GNASXL*, whose out-group species is the squirrel monkey (this dataset consists of primates only). The results of LRT analyses between models M1, M2, M7, M8, and M8a are given in the last column. An agreement across all models is required for positive selection to be detected. Based on this criterion, the symbol "+" represents cases where positive selection was found and "−" represents cases where there is no evidence of positive selection

functional links to the evolutionarily conserved sterol binding pocket that has recently been reported for this class of proteins, but find that the residues under positive selection fall outside of the known functional residues (Suchanek et al. 2007; Wang et al. 2008).

The signature of positive selection detected for the *GNASXL* is considered to be due to a scenario of evolutionary oscillation resulting from its protein:protein interaction with the product of the overlapping gene *ALEX* (Wadhawan et al. 2008). We also tested whether *ALEX* displayed site-specific and lineage-specific evolution similar to *GNASXL*, and found using site-specific models that the *ALEX* locus is also under positive Darwinian selection (Results not shown). While it was possible to detect rapidly evolving codons for both *GNASXL* and *ALEX*, the signature of positive selection detected for *GNASXL* and *ALEX* may not necessarily be linked to the imprinted status of *GNASXL* (Wadhawan et al. 2008).

In mammals, there is no correlation between imprinting status and evidence of lineage-specific positive Darwinian selection.

For each of the 34 mammalian imprinted genes, we also tested for evidence of positive selection on branches leading to placental mammal species where the imprinting status of the particular gene is known. To determine if there has been variation in evolutionary rate following the advent (or loss) of imprinting within a clade, and specifically whether there is any evidence of positive selection we analysed the 34 multiple sequence alignments using lineage and site-specific models of evolution. The results indicate that for 14 genes there is evidence of positive selection in some orthologs of imprinted loci, namely: *PLAGL1* (mouse lineage); *IGF2* (mouse, opossum); *SLC22A18* (mouse); *OSBPL5* (mouse, cow); *DCN* (mouse); *DLK1* (mouse, human); *RASGRF1* (mouse, rat); *IGF2R* (opossum, rat); *IMPACT* (mouse); *GRB10* (human); *NAP1L4* (mouse); *UBE3A* (mouse); *GATM* (mouse); *GRBRG3* (human) (Fig. 3; Table 2).

However our lineage-specific analysis of selection also demonstrates that there are 33 cases where the ortholog is imprinted for that lineage but there is no evidence of positive selection (Fig. 3; Table 2). There are also examples of orthologs of imprinted genes in certain species that are known to be biallelically expressed (e.g. *OSPBL5* in cow), yet these genes display evidence of positive selection (Fig. 3; Table 3). Therefore, there is a lack of correlation between imprinting status and evidence of lineage-specific positive selection.

Orthologs of imprinted genes in mammalian lineages display high levels of positive Darwinian selection, irrespective of their known imprinting status.

The imprinted status of many of the orthologs of the known imprinted genes across different mammalian lineages unfortunately remains unknown. Nonetheless, we also screened for evidence of positive selection across all lineages irrespective of whether the imprint status of the ortholog was known or not in the particular lineage

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

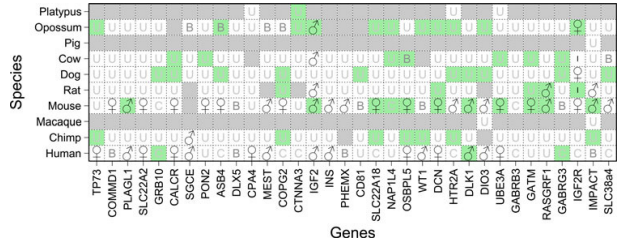


Fig. 3 Summary of positive selection analysis on mammalian imprinted genes. The "imprinted clade" orthologs are listed on the x-axis (*left* to *right*). The species tested are listed on the y-axis, from human to platypus (the outgroups chicken and fish are excluded from the Table). In each box the known imprinting status of the gene in the lineage is indicated i.e.: *male sign* paternally expressed imprinted,; *female sign* maternally expressed imprinted, *U* unknown imprinted status, *B* biallelically expressed, *C* conflicting data. *Green colored boxes* indicate those lineages where positive selection was found for a particular gene. *White boxes* indicate that there was no evidence of positive selection detected in the particular lineage. *Gray boxes* are cases with no data i.e. where no homolog could be found for the dataset

(Table 3). Our analysis of multiple sequence alignments for 34 different imprinted gene loci (imprinted in one or more lineages) indicate that, out of 158 lineages of unknown imprinting status tested, 36 lineages display evidence of lineage specific positive selection (Fig. 3) across a total of 29 orthologous genes; *PLAGL1; TP73; GRB10; CALCR; PON; ASB4; COPG2; CTNNA3; CD81; SLC22a18; NAP1L4; OSBPL5; WT1; DCN; HRT2A; DLK1; DIO3; UBE3A; GATM; GABRG3; IMPACT; SLC38a4; IGF2; SLC22A18; OSBPL5; DCN; RASGRF1; IGF2R* and *IMPACT*.

Overall our results indicate that many of the loci prone to epigenetic regulation by imprinting in mammals are also subject to positive selection in different lineages. However, the positive selection detected is independent of known imprint status (i.e. whether they are imprinted or biallelic). The lineage specific analyses are summarised in Fig. 2.

Discussion

The relationship between the molecular evolution of imprinted loci and their imprinted status remains unclear for most imprinted genes, particularly in relation to evidence supporting different theories regarding the evolution of imprinted loci (Hurst 1997; Wood and Oakey 2006). The rapid growth in genome sequencing data is now facilitating comparison of imprinted loci across mammalian species and outgroup species, with a range of functional studies indicating that imprinting control mechanisms are not highly conserved across species (Killian et al. 2000; Lawton et al. 2008; Vu et al. 2006; Weidman et al. 2006a; Weidman et al. 2006b; Weidman et al. 2004). For instance, differences in imprinting mechanisms (e.g. location of differentially methylated regions (DMRs), DMR activation vs. repression, absence vs. presence of non-coding RNAs) are observed between eutherians and metatherians for imprinted genes such as *IGF2* and *IGF2R* (Lawton et al. 2008; Weidman et al. 2006a). This could suggest that orthologs of imprinted genes could be undergoing different selection pressures in different lineages where imprinting is known to be possible. In this study, we undertook a phylogenetic approach to investigate whether the molecular evolution of known imprinted protein-coding genes provided support for parental conflict (kinship) and/or dosage compensation theories for the evolution of genomic imprinting.

While parental conflict theories for the evolution of imprinted genes are widely considered, there is a paucity of evidence at the molecular level to support parental conflicts as a major driver of the evolution of most imprinted loci. In this study, using the orthologous genes located within the clade that contains imprinted genes (Fig. 1), we investigated whether any imprinted protein-coding genes displayed evidence for positive selection (Figs. 2 and 3; Table 1). We also determined whether there was any obvious correlation between the imprinting status of an ortholog and the selective pressure (i.e. purifying vs positive selection) observed in any given lineage (Fig. 3; Tables 2 and 3).

In a recent molecular evoutionary comparison of Ka/Ks values (akin to Dn/Ds) for mouse-human gene pairs of imprinted loci, it was found that the imprinted genes *Cdkn1c*, *Phlda2* and *Usp29* had the highest Ka/Ks values, albeit all under 0.5 which is consistent with purifying selection (Hutter et al. 2010). In that study, it was highlighted that the paternally expressed imprinted genes *Snrpn*, *Mest*, *Wt1* and *Copg2* were all under strong purifying selection. Overall, the study concluded that the set of

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

Table 2 Lineage specific positive Darwinian selection for mammalian imprinted genes for which the imprinted status of the gene is known

Gene	Mammalian species for which imprinting status is known	Lineage specific positive selection
TP73	Human ($+^{m}$)	–
COMMD1	Human (–), mouse ($+^{m}$)	–/–
PLAGL1	Human ($+^{p}$), *mouse* ($+^{p}$)	–/+ (1% of sites, $\omega = 999$)
SLC22A2	Human ($+^{m}$), mouse ($+^{m}$),	–/–
GRB10	*Human* ($+^{m}$ *and* $+^{p}$), mouse ($+^{m}$ and $+^{p}$)	+(16% of stes, $\omega = 26.23$)/–
CALCR	Human ($+^{m}$), mouse ($+^{m}$)	–/–
SGCE	Human ($+^{p}$), mouse ($+^{p}$), opossum (–)	–/–/–
PON2	Mouse ($+^{m}$)	–
ASB4	Human (CD), mouse ($+^{m}$), *opossum* (–)	–/–/+ (34% of sites, $\omega = 4.15$)
DLX5	Human (–), mouse (–)	–/–
CPA4	Human ($+^{m}$)	–
MEST (Peg1)	Human ($+^{p}$), mouse ($+^{p}$), opossum (–)	–/–/–
COPG2	Human (CD), mouse ($+^{m}$), opossum (–)	–/–/–
CTNNA3	Human (CD)	–
IGF2	Cow ($+^{p}$), *opossum* ($+^{p}$), *mouse* ($+^{p}$), rat ($+^{p}$)[a], human ($+^{p}$)	–/+ (20% of sites, $\omega = 28.25$)/+ (3% of sites, $\omega = 999$)/–/–
INS	Human ($+^{p}$), mouse ($+^{p}$)	–/–
PHEMX	Human (–), mouse ($+^{m}$)	–/–
CD81	Mouse (–)	–
SLC22A18	Human (CD), *mouse* ($+^{m}$)	–/+ (2% of sites, $\omega = 23.22$)
NAP1L4	*Mouse* (*CD*)	+ (1% of sites, $\omega = 56.86$)
OSBPL5	Human ($+^{m}$)[b], *mouse* ($+^{m}$), *cow* (–)	–/+ (2% sites, $\omega = 94.02$)/+(2% sites, $\omega = 999$)
WT1	Human ($+^{p}$), mouse (–)	–/–
DCN	Human ($+^{m}$), *mouse* ($+^{m}$)	–/+ (7% of sites, $\omega = 3.77$)
HTR2A	Human (CD), mouse ($+^{m}$)	–/–
DLK1	*Human* ($+^{p}$), *mouse* ($+^{p}$)	+ (3% of sites, $\omega = 999$)/+ (17% of sites, $\omega = 999$)
DIO3	Human ($+^{p}$); mouse ($+^{p}$)	–/–
UBE3A	Human ($+^{m}$), *mouse* ($+^{m}$)	–/+ (13% of stes, $\omega = 1.107$)
GABRB3	Human (CD), mouse (CD)	–/–
GATM	*Mouse* ($+^{m}$)	+(51% of stes, $\omega = 8.685$)
RASGRF1	*Mouse* ($+^{p}$), *rat* ($+^{p}$)	+ (8% of sites, $\omega = 4.84$)/+ (5% of sites, $\omega = 4.93$)
GABRG3	*Human* (*CD*), mouse (–)	+(0.5% of stes, $\omega = 177$) / -
IGF2R	Human (CD), mouse ($+^{m}$), dog ($+^{m}$), cow (+), *opossum* ($+^{m}$), *rat* (+)	–/–/–/–/+ (3% of sites, $\omega = 1.30$)/+ (1% of sites, $\omega = 6.77$)
IMPACT	Human (–), *mouse* ($+^{p}$), rat ($+^{p}$)	–/+ (4% of sites, $\omega = 16.35$)/–
SLC38a4	Human (CD), mouse ($+^{p}$), cow (–)	–/–/–
Total	75 Lineages tested	20 With positive selection

CD conflicting data, $+^{m}$ maternally expressed imprinted, $+^{p}$ paternally expressed imprinted. (–) biallelic expression, (+) imprinted of unknown parent-of-origin, [a] paternally expressed in most tissues except choroid plexus and meninges, [b] result for OSBPL5 is unreliable for the prediction of positively selected sites in this lineage, *NP* not present. [c] Including OSBPL5 Human. The *left* column represents the genes in the dataset. The *central* column represents those species for which a homolog of the gene exists and whose imprinting status is currently known. Species names in *italics* represent positive selection detected in those lineages. The values in the column labelled "*lineage specific positive selection*" are all the results of the selection analysis for each of the species. The lineage specific positive selection results are listed in the *same order* as the species appear in the table. The symbol "–" marks those cases where no positive selection is found and "+" marks those cases where it has been detected. The values immediately following the "+" cases of positive selection are the proportions of sites (given as a percentage of the alignment length) in that lineage that have an ω value greater than 1 (varies from gene to gene and lineage to lineage)

imprinted genes studied did not contain genes with exceptionally high Ka/Ks ratios that could be indicative of recent positive selection (Hutter et al. 2010).

In contrast, our analysis of the molecular evolution of the 34 mammalian imprinted genes identifies two imprinted genes which are subject to site specific positive

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

Author Proof

Table 3 Analysis of lineage-site specific evolution on the orthologous genes of unknown imprinting status using lineage specific model of evolution

Gene	Mammalian species of unknown imprinting status	Lineage specific positive selection
TP73	*Chimp*, cow, dog, mouse, *opossum*, rat	+ (24% of sites, $\omega = 160.84$)/–/–/–/+ (19% of sites, $\omega = 26.32$)/–
COMMD1	Chimp, cow, dog, opossum, rat	–/–/–/–/–
PLAGL1	Chimp, cow, dog, opossum, rat	–/–/–/–/–
SLC22A2	Chimp, cow, dog, opossum, rat	–/–/–/–/–
GRB10	Chimp, cow, *dog*, opossum, rat	–/–/+ (16% of sites, $\omega = 26.23$)/–/–
CALCR	Chimp, *cow*, *dog*, rat	–/+ (12% of sites, $\omega = 1.14$)/+ (2% of sites, $\omega = 60.70$)/–
SGCE	Chimp, cow, dog	–/–/–
PON2	Chimp, *cow*, dog, human, opossum, rat	–/+ (8% of sites, $\omega = 7.54$)/–/–/–/–
ASB4	Chimp, cow, *dog*, rat	–/–/+ (0.25% of sites, $\omega = 999$)/–
DLX5	Chimp, cow, dog, opossum, rat	–/–/–/–/–
CPA4	Chimp, dog, mouse, opossum, platypus, rat	–/–/–/–/–/–
MEST	Chimp, cow, dog	–/–/–
COPG2	*Chimp*, cow, *dog*, *rat*	+ (2% of sites, $\omega = 936.11$)/–/+ (2% of sites, $\omega = 9.07$)/+ (0.3% of sites, $\omega = 600$)
CTNNA3	Chimp, cow, dog, mouse, *opossum*, *platypus*	–/–/–/–+ (2% of sites, $\omega = 214$)/+ (9% of sites, $\omega = 138$)
IGF2	Chimp, dog	–/–
INS	Chimp, cow, dog, rat	–/–/–/–
PHEMX	Cow, dog, rat	–/–/–
CD81	Chimp, cow, *dog*, human, rat	–/–/+ (27% of sites, $\omega = 383$)/–/–
SLC22A18	*Chimp*, cow, dog, *opossum*, rat	+ (2% of sites, $\omega = 999$)/–/–/+ (23% of sites, $\omega = 337$)/–
NAP1L4	Chimp, *cow*, dog, human, *opossum*, rat	–/+ (2% of sites, $\omega = 112$)/–/–/+ (4% of sites, $\omega = 15.11$)/–
OSBPL5	*Chimp*, dog, opossum, rat	+ (1% of sites, $\omega = 999$)/–/–/–
WT1	*Chimp*, dog, *opossum*, rat	+ (12% of sites, $\omega = 999$)/–/+ (6% of sites, $\omega = 899$)/–
DCN	Chimp, cow, dog, *opossum*, *rat*	–/–/–/+ (4% of sites, $\omega = 998$)/+ (9% of sites, $\omega = 3.5$)
HTR2A	*Chimp*, cow, *dog*, *opossum*, platypus, rat	+ (1% of sites, $\omega = 164$)/–/+ (2% of sites, $\omega = 24$)/+ (2% of sites, $\omega = 998$)/–/–
DLK1	Chimp, cow, *dog*, *opossum*, rat	–/–/+ (3% of sites, $\omega = 999$)/+ (4% of sites, $\omega = 20.54$)/–
DIO3	Macaque, cow, *dog*, *opossum*	–/–/+ (6% of sites, $\omega = 4.81$)/+ (16% of sites, $\omega = 1.03$)
UBE3A	Chimp, *cow*, dog, opossum, platypus, rat	–/+ (2% of sites, $\omega = 6.83$)/–/–/–/–/–
GABRB3	Chimp, cow, dog, opossum, rat	–/–/–/–/–
GATM	Chimp, *cow*, dog, human, opossum, *rat*	–/+ (1% of sites, $\omega = 34.40$)/–/–/–/+ (3% of sites, $\omega = 612$)
RASGRF1	Chimp, cow, dog, human, opossum	–/–/–/–/–
GABRG3	Chimp, *cow*, *dog*, opossum, rat	–/+ (2% of sites, $\omega = 998$)/+ (1% of sites, $\omega = 177$)/–/–
IGF2R	Chimp	–
IMPACT	*Chimp*, pig, macaque, cow, dog, opossum	+ (3% of sites, $\omega = 999$)/–/–/–/–/–
SLC38a4	Chimp, *dog*, opossum, rat	–/+ (1% of sites, $\omega = 26.22$)/–/–
Total	158 Lineages tested	36 with positive selection

The *left* column represents the genes in the dataset. The *central* column represents those species for which a homolog of the gene exists but whose imprinting status is currently unknown. Species names in *italics* represent positive selection detected in those lineages. The values in the column labeled "*lineage specific positive selection*" are all the results of the selection analysis for each of the species of unknown imprinting status. The lineage specific positive selection results are listed in the *same order* as the species appear in the table. The symbol "–" marks those cases where no positive selection is found and "+" marks those cases where it has been detected. The values immediately following the "+" cases of positive selection are the proportions of sites (given as a percentage of the alignment length) in that lineage that have an ω value greater than 1 (varies from gene to gene and lineage to lineage)

UNCORRECTED PROOF

selection (*OSBPL5*, *GNASXL*) (Table 1), and identifies a number of genes which are both imprinted and undergoing positive selection in specific (recent) lineages (i.e. *PLAGL1*, *IGF2*, *SLC22A18*, *OSBPL5*, *DCN*, *DLK1*, *RASGRF1*, *IGF2R* and *IMPACT*) (Fig. 3; Table 2). While the detection of two codons under positive selection in the imprinted gene *OSBPL5* [a sterol binding protein (Suchanek et al. 2007)] is intriguing, no functions for these

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

codons in *OSBPL5* has yet been reported. In the case of *GNASXL* it has previously been shown that *GNASXL* is rapidly evolving, possibly due to its overlapping reading frame with *ALEX* (Wadhawan et al. 2008). Our results confirmed the rapid evolution of both *GNASXL* and *ALEX*. While the mutual binding affinity between the GNASXL and ALEX proteins can maintain an oscillating evolutionary scenario, it remains unclear what mechanism would have initiated one or other of these two genes to become subject to positive selection at specific sites.

At present, a comprehensive assessment of the extent of functional conservation of imprinting (both status and mechanism) across placental animal species is not yet available. However, using the available imprinting status data, our results do indicate an overall lack of concordance between known imprinting status and evidence of positive selection (Figs. 2 and 3; Table 2). It remains possible that a subset of imprinted protein-coding genes (i.e. *PLAGL1, IGF2, SLC22A18, OSBPL5, DCN, DLK1, RASGRF1, IGF2R* and *IMPACT)* in some lineages are involved in antagonistic co-evolution (with other factors regarding resource allocation to the offspring) via a parental conflict mediated by amino-acid changes in proteins encoded by imprinted loci (Figs. 2 and 3; Table 2). However, given that we can detect orthologs of imprinted genes which are biallelically expressed and under positive selection in the same lineage, we consider that the evidence found for positive selection of known imprinted genes in specific lineages could as likely be due to neo-functionalisation processes, rather than rapid-evolution of proteins driven by antagonistic co-evolution. We recognize that imprinted genes can display tissue-specific imprinting and that cases of tissue or strain specific imprinting may yet be detected in lineages where it is currently considered that the imprinted gene ortholog is biallelically expressed. However, only two of the fourteen cases that are considered as biallelically expressed in this study displayed evidence of positive selection (Fig. 3), suggesting that changes of imprinting status for any or all of the biallelically expressed genes would not significantly alter our findings. We also detected significant evidence of lineage-specific positive (and purifying) selection for orthologs of imprinted genes in lineages where it is not yet known whether the ortholog is imprinted or not (Fig. 3 and Table 3). Overall, our results do not provide generic support for any majority of imprinted protein-coding loci being involved in antagonistic co-evolution via a parental conflict which is mediated by amino-acid changes in proteins encoded by imprinted loci.

In contrast to parental conflict theories, dosage compensation theories for the evolution of imprinting are based on an association between gene (or genome) duplication and gene dosage reduction via genomic imprinting (Iwasa 1998; Okamura et al. 2004; Walter and Paulsen 2003; Wood and Oakey 2006). Our analysis of the extent of gene duplication history for imprinted genes within the imprinted clade suggests that the majority of imprinted genes have not undergone ancestral *trans*- or *cis*-duplications within the imprinting clade (i.e. the clade which includes orthologs of imprinted genes from platypus, opossum, pig, cow, dog, rat, mouse, macaque, chimp and human). A previous analysis of 21,115 human protein coding genes excluded recently duplicated gene families and identified 16,529 conserved 1:1 ortholog sets across six complete mammalian genomes (Kosiol et al. 2008). The 1:1 ortholog sets identified contained a human gene and either five (42% of the 16,529 cases), four (28%), three (15%) or two (15%) non-human orthologs. This indicates that ~33% of the 21,115 genes are retained as 1:1 orthologs across at least five mammalian genomes. With the caveat that the test imprintome in this study of 34 genes is an extremely small sample size, we observe that 88% (30) of the imprinted loci analysed are retained as single gene 1:1 orthologs across five to nine mammalian species. The kinship theory predicts that an equilibrium state would occur whereby expression of paternally expressed imprinted genes would be at a paternal optimum, while expression of maternally expressed imprinted genes would be at a maternal optimum. Hence, it could be expected that increases (or changes) in expression of imprinted loci due to duplications, biallelic expression or uniparental disomy would be selected against (Haig 2006; Wilkins and Haig 2003). Our results suggest that imprinted loci may be less prone to duplications than other genes in the human genome, which remains compatible with both the kinship and dosage hypotheses. The overall lack of retained gene duplicates in the imprinted gene set was further reinforced by the strong micro-synteny conservation observed for the majority of imprinted loci examined (*Supplementary File 4*).

Overall, our results indicate that following the emergence of imprinting within the placental mammals, subsequent gene duplication and retention events within the "imprinted clade" are rare. The presence of imprinted genes predominantly as singletons within the imprinted clade could be suggestive of a model involving selection against *cis*- or *trans*-duplication of these genes, possibly as a means of limiting the dosage of these genes. Imprinting may have arisen as a "dosage-defense" mechanism for limiting expression levels of duplicated genes (e.g. *IMPACT*, *IGF2* in mammals and *MEDEA* in plants) within gene regulatory networks that display dosage-sensitive effects on the organism (Iwasa 1998; Okamura et al. 2004; Varrault et al. 2006; Walter and Paulsen 2003; Wood and Oakey 2006).

While our results provide little evidence of the protein coding regions of imprinted genes evolving under positive selection, it remains possible that the regulatory regions of imprinted loci could be fast evolving under positive selection in a manner that affects the timing, location or level of gene expression. Indeed, the parental conflict and dosage compensation theories within specific lineages may not be mutually exclusive as the conflict theory could also be consistent with evolutionary change resulting from gene duplication and losses that would provide a mechanism to overcome maternal or paternal control that operates faster than the process of point mutations. Our results highlight that the true orthologs of imprinted genes display little evidence of *cis*- or *trans*-duplications and are evolving differentially across lineages, in a manner whereby there is no obvious correlation between the imprinted status of the ortholog in the lineage and whether the protein coding region of the imprinted locus is under positive or purifying selection.

Acknowledgements CS & MJO'C acknowledge funding support from Science Foundation Ireland. MD & NBL acknowledge funding support from IRCSET. TAW acknowledges funding support from the School of Biotechnology and the Pierse Trust DCU. We would like to thank the Science Foundation Ireland and the Higher Education Authority - Irish Centre for High-End Computing (ICHEC) for processor time and technical support. We are grateful to the comments of two anonymous reviewers on a previous draft of this paper. We would like to thank Dr James McInerney (NUIM, Ireland) and Prof Ken Wolfe (TCD, Ireland) for discussions relating to analysis and for computational facilities. CS and MJOC are funded by Science Foundation Ireland (Grants 08/IN.1/B1931 and EOB2673).

Conflict of interest The authors declare no conflict of interest.

References

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5:e1000262

Chang BS, Campbell DL (2000) Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. Mol Biol Evol 17:1220–1231

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31:3497–3500

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113

Foster PG, Jermiin LS, Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol 44:282–288

Gamier O, Laoueille-Duprat S, Spillane C (2008) Genomic imprinting in plants. Epigenetics 3:14–20

Glaser RL, Ramsay JP, Morison IM (2006) The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. Nucleic Acids Res 34:D29–D31

Greally JM (2002) Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. Proc Natl Acad Sci USA 99:327–332

Haig D (2004) Genomic imprinting and kinship: how good is the evidence? Annu Rev Genet 38:553–585

Haig D (2006) Intragenomic politics. Cytogenet Genome Res 113: 68–74

Haig D, Trivers R (1995) The evolution of parental imprinting: a review of hypotheses. In: Ohlsson R, Hall K, Ritzen M (eds) Genomic imprinting: causes and consequences. Cambridge University Press, Cambridge, pp 17–28

Hurst LD (1997) Evolutionary theories of genomic imprinting. In: Reik W, Surani A (eds) Genomic imprinting. IRL Press, Oxford, pp 211–237

Hurst LD, McVean GT (1998) Do we understand the evolution of genomic imprinting? Curr Opin Genet Dev 8:701–708

Hutter B, Bieg M, Helms V, Paulsen M (2010) Divergence of imprinted genes during mammalian evolution. BMC Evol Biol 10:116

Inagaki Y, Roger AJ (2006) Phylogenetic estimation under codon models can be biased by codon usage heterogeneity. Mol Phylogenet Evol 40:428–434

Iwasa Y (1998) The conflict theory of genomic imprinting: how much can be explained? Curr Top Dev Biol 40:255–293

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Keane TM, Naughton TJ, McInerney JO (2007) MultiPhyl: a high-throughput phylogenomics webserver using distributed computing. Nucleic Acids Res 35:W33–W37

Killian JK, Byrd JC, Jirtle JV, Munday BL, Stoskopf MK, MacDonald RG, Jirtle RL (2000) M6P/IGF2R imprinting evolution in mammals. Mol Cell 5:707–716

Kono T (2006) Genomic imprinting is a barrier to parthenogenesis in mammals. Cytogenet Genome Res 113:31–35

Kono T, Obata Y, Wu Q, Niwa K, Ono Y, Yamamoto Y, Park ES, Seo JS, Ogawa H (2004) Birth of parthenogenetic mice that can develop to adulthood. Nature 428:860–864

Kono T, Kawahara M, Wu Q, Hiura H, Obata Y (2006) Paternal dual barrier by Ifg2–H19 and Dlk1–Gtl2 to parthenogenesis in mice. Ernst Schering Res Found Workshop, pp 23–33

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six mammalian genomes. PLoS Genet 4:e1000144

Lawton BR, Carone BR, Obergfell CJ, Ferreri GC, Gondolphi CM, Vandeberg JL, Imumorin I, O'Neill RJ, O'Neill MJ (2008) Genomic imprinting of IGF2 in marsupials is methylation dependent. BMC Genomics 9:205

Lercher MJ, Hurst LD (2003) Imprinted chromosomal regions of the human genome have unusually high recombination rates. Genetics 165:1629–1632

Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, Hartemink AJ (2007) Computational and experimental identification of novel human imprinted genes. Genome Res 17:1723–1730

McVean GT, Hurst LD (1997) Molecular evolution of imprinted genes: no evidence for antagonistic coevolution. Proc Biol Sci 264:739–746

Moore T, Haig D (1991) Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet 7:45–49

O'Neill MJ, Lawton BR, Mateos M, Carone DM, Ferreri GC, Hrbek T, Meredith RW, Reznick DN, O'Neill RJ (2007) Ancient and continuing Darwinian selection on insulin-like growth factor II in placental fishes. Proc Natl Acad Sci USA 104:12404–12409

Okamura K, Yamada Y, Sakaki Y, Ito T (2004) An evolutionary scenario for genomic imprinting of Impact lying between nonimprinted neighbors. DNA Res 11:381–390

Page RD (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14:819–820

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | LE | TYPESET
MS Code : MG-OC-10-0086 | CP | DISK

Pardo-Manuel de Villena F, de la Casa-Esperon E, Sapienza C (2000) Natural selection and the function of genome imprinting: beyond the silenced minority. Trends Genet 16:573–579

Parker-Katiraee L, Carson AR, Yamada T, Arnaud P, Feil R, Abu-Amero SN, Moore GE, Kaneda M, Perry GH, Stone AC, Lee C, Meguro-Horike M, Sasaki H, Kobayashi K, Nakabayashi K, Scherer SW (2007) Identification of the imprinted KLF14 transcription factor undergoing human-specific accelerated evolution. PLoS Genet 3:e65

Pask AJ, Papenfuss AT, Ager EI, McColl KA, Speed TP, Renfree MB (2009) Analysis of the platypus genome suggests a transposon origin for mammalian imprinting. Genome Biol 10:R1

Sandovici I, Kassovska-Bratinova S, Vaughan JE, Stewart R, Leppert M, Sapienza C (2006) Human imprinted chromosomal regions are historical hot-spots of recombination. PLoS Genet 2:e101

Schmidt HA, von Haeseler A (2007) Maximum-likelihood analysis using TREE-PUZZLE. Curr Protoc Bioinformatics, Chap 6, Unit 6 6

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Smith NG, Hurst LD (1998) Molecular evolution of an imprinted gene: repeatability of patterns of evolution within the mammalian insulin-like growth factor type II receptor. Genetics 150: 823–833

Spillane C, Schmid KJ, Laoueille-Duprat S, Pien S, Escobar-Restrepo JM, Baroux C, Gagliardini V, Page DR, Wolfe KH, Grossniklaus U (2007) Positive darwinian selection at the imprinted MEDEA locus in plants. Nature 448:349–352

Suchanek M, Hynynen R, Wohlfahrt G, Lehto M, Johansson M, Saarinen H, Radzikowska A, Thiele C, Olkkonen VM (2007) The mammalian oxysterol-binding protein-related proteins (ORPs) bind 25-hydroxycholesterol in an evolutionarily conserved pocket. Biochem J 405:473–480

Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20:18–20

Varmuza S, Mann M (1994) Genomic imprinting-defusing the ovarian time bomb. Trends Genet 10:118–123

Varrault A, Gueydan C, Delalbre A, Bellmann A, Houssami S, Aknin C, Severac D, Chotard L, Kahli M, Le Digarcher A, Pavlidis P, Journot L (2006) Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth. Dev Cell 11:711–722

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19: 327–335

Vu TH, Jirtle RL, Hoffman AR (2006) Cross-species clues of an epigenetic imprinting regulatory code for the IGF2R gene. Cytogenet Genome Res 113:202–208

Wadhawan S, Dickins B, Nekrutenko A (2008) Wheels within wheels: clues to the evolution of the Gnas and Gnal loci. Mol Biol Evol 25:2745–2757

Walter J, Paulsen M (2003) The potential role of gene duplications in the evolution of imprinting mechanisms. Hum Mol Genet 12(Spec no 2):R215–R220

Wang PY, Weng J, Lee S, Anderson RG (2008) The N terminus controls sterol binding while the C terminus regulates the scaffolding function of OSBP. J Biol Chem 283:8034–8045

Weidman JR, Murphy SK, Nolan CM, Dietrich FS, Jirtle RL (2004) Phylogenetic footprint analysis of IGF2 in extant mammals. Genome Res 14:1726–1732

Weidman JR, Dolinoy DC, Maloney KA, Cheng JF, Jirtle RL (2006a) Imprinting of opossum Igf2r in the absence of differential methylation and air. Epigenetics 1:49–54

Weidman JR, Maloney KA, Jirtle RL (2006b) Comparative phylogenetic analysis reveals multiple non-imprinted isoforms of opossum Dlk1. Mamm Genome 17:157–167

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Wilkins JF, Haig D (2003) What good is genomic imprinting: the function of parent-specific gene expression. Nat Rev Genet 4: 359–368

Wood AJ, Oakey RJ (2006) Genomic imprinting in mammals: emerging themes and established theories. PLoS Genet 2:e147

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409–418

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118

Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472–2479

Journal : Large 335 | Dispatch : 15-9-2010 | Pages : 13
Article No. : 9283 | ☐ LE | ☐ TYPESET
MS Code : MG-OC-10-0086 | ☑ CP | ☑ DISK

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins

Claire C Morgan, Noeleen B Loughran, Thomas A Walsh, Alan J Harrison, Mary J O'Connell*

Abstract

Background: Reproductive proteins are central to the continuation of all mammalian species. The evolution of these proteins has been greatly influenced by environmental pressures induced by pathogens, rival sperm, sexual selection and sexual conflict. Positive selection has been demonstrated in many of these proteins with particular focus on primate lineages. However, the *mammalia* are a diverse group in terms of mating habits, population sizes and germ line generation times. We have examined the selective pressures at work on a number of novel reproductive proteins across a wide variety of *mammalia*.

Results: We show that selective pressures on reproductive proteins are highly varied. Of the 10 genes analyzed in detail, all contain signatures of positive selection either across specific sites or in specific lineages or a combination of both. Our analysis of SP56 and Col1a1 are entirely novel and the results show positively selected sites present in each gene. Our findings for the Col1a1 gene are suggestive of a link between positive selection and severe disease type. We find evidence in our dataset to suggest that interacting proteins are evolving in symphony: most likely to maintain interacting functionality.

Conclusion: Our *in silico* analyses show positively selected sites are occurring near catalytically important regions suggesting selective pressure to maximize efficient fertilization. In those cases where a mechanism of protein function is not fully understood, the sites presented here represent ideal candidates for mutational study. This work has highlighted the widespread rate heterogeneity in mutational rates across the *mammalia* and specifically has shown that the evolution of reproductive proteins is highly varied depending on the species and interacting partners. We have shown that positive selection and disease are closely linked in the Col1a1 gene.

Background

Reproductive proteins are essential for success of sexually reproducing species and indeed for the emergence of new species. In the past it has been observed that reproductive proteins tend to be under positive selective pressure to change, i.e. adaptive evolution, a trend found in a variety of animal species from abalone to primates [1,2]. Adaptive evolution or positive selection is a selective pressure placed on a protein by a change in environment in order to improve the fitness of the organism in that environment.

With changes in environment, that can include mating system, there is a subsequent selective pressure on the protein sequences related to those functions to adapt accordingly. This variation can be detected using the well-known measurements of the rate of non-synonymous substitutions per non-synonymous site (Dn) and synonymous substitutions per synonymous site (Ds) and their ratio ω = Dn/Ds. The detection of adaptive evolution, where the ratio exceeds unity, is referred to as positive Darwinian selection. Detecting positive Darwinian selection in a region of a protein, or indeed in a lineage of a phylogeny, indicates that there is a selective advantage in changing the amino acid sequence in this region. These signals are essential for our understanding of functionally important residues in a protein sequence and protein functional shift.

In general, the rate of mutation that a gene undergoes is contingent on a number of factors including; protein structure, presence of gene duplicates, location in the genome, effective population size, germ line generation time, and composition of the sequence (for review see [3]). It has recently been shown that the number of physical interactions of a particular protein also influences the intrinsic rate of evolution [4]. Evidence for the generation time effect has come from studies on various proteins and species including analyses of substitution rates in higher primates and rodents [5], substitution rates in higher grasses and in palms [6], in mammalian genomes [7] and in chloroplast and sex mutation rate ratios [5,6]. With recent advances in sequencing we have an opportunity to examine these effects using a wider selection of proteins and species. Documented selective pressures associated with positive selection in reproductive proteins include: (i) intense sperm competition whereby sperm from numerous males, ejaculated into the female reproductive tract, compete with one another for the prized fertilization of the egg [8]; (ii) evasion of the immune system, whereby surface layer reproductive proteins evolve to evade destruction by the host's immune system [8]; and finally (iii) selective pressures enforced by mating system, related of course to point (i) above. Species that are more promiscuous have increased levels of selective pressure acting on reproductive proteins than species that are monogamous. This later point is illustrated in the study of SEMG2, where adaptive evolution was found to correlate with mating system in primates [9].

In order to determine the variation in selective pressure in these proteins, there are a number of criteria that the data must meet. Firstly, the data must have a robust phylogenetic signal. Secondly, systematic biases that may exist in the data must be minimized, these include but are not limited to: long branch attraction (LBA), amino acid composition bias, base composition bias and unqualified ortholog predictions, all of which may lead to inaccurate estimates of phylogeny. Thirdly, sensitivity to taxa number is a known limitation of methods for detecting positive selection, therefore more than 6 taxa are needed to gain accurate estimations of selective pressure using the maximum likelihood (ML) method applied here [10].

In this study we have selected a subset of proteins that have roles to play in reproduction. Our dataset was composed of three major datatypes, (i) previously published reproductive proteins, (ii) interacting proteins, here we identified proteins shown to interact with (i), and finally (iii), genes identified from microarray experiments as being highly expressed in reproductive tissues. For group (iii) we assume that those proteins highly expressed in reproductive tissues are important for the function of that tissue. The previously untested reproductive proteins analysed here are from data types (ii) and (iii) outlined above. These novel proteins are SP56, Porimin and Col1a1. SP56 is sperm binding protein number 56, this protein is a representative of the interacting protein subset of sequences analysed. SP56 has been shown to interact with ZP3 - a well-studied reproductive protein. Both Porimin and Col1a1 have been identified from published microarray experiments on normal human tissue [11], and were selected for analysis due to their high levels of expression in reproductive tissues in that study. Porimin is a transmembrane protein that is highly expressed in the uterus, prostate and placenta and Col1a1 is highly expressed in the uterus. Further evidence for the link between Porimin and reproduction was not available in the literature and therefore results from this particular gene are taken with caution until this protein is further characterized. Col1a1 plays an important role during spermatogenesis where it mediates the detachment and migration of germ cells, thus adding further support for its role in reproduction [12].

We have analyzed these data with an approach sensitive to all the systematic biases and limitations of methods given above. A number of genes in our dataset have been analyzed previously but have not taken these limitations and considerations into account. We have expanded these datasets to include a greater number of taxa, we have analyzed all of these genes for evidence of systematic biases and we have used improved models of codon evolution. In this paper we have included models that allow for rate variation across the sequence and across the phylogeny.

Results and Discussion

We performed phylogenetic analyses on all 11 datasets. The resultant gene trees were found to conflict with the canonical phylogeny species ([13], as adapted in Figure 1. The only exception was the Catsper1 mammalian dataset. We postulate the following causes for this conflict: (1) amino acid and/or base composition bias, (2) lack of phylogenetic signal in the data, and finally (3), LBA caused by mixtures of long and short germ line generation times (see Figure 2 for a sample of species and their germ line generation times from our dataset). What follows is a summary of the results of the tests of data quality and bias we performed, see Table 1 for synopsis. We carried out these tests to determine in each case whether these conflicting phylogenies are accurate descriptions of history or whether the data are subject to these known issues listed 1-3 above. Subsequent statistical comparison of the gene trees and species phylogeny using the Shimodaira Hasegawa (SH) test [14] revealed that there is no statistical difference between the gene and species trees in each case, see Table 2 for results of SH tests. The only exceptions

* Correspondence: mary.oconnell@dcu.ie
Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland

Table 1 Summary of the analysis of quality and bias present in the data

GENE	DATA QUALITY			PHYLOGENETIC ANALYSIS		
	LM Category	AA Comp Bias	Base Comp Bias	Substitution Model	Gene v Species Tree	LBA Artifact
Adam2	1	Pass	Pass	JTT+G	Unresolved	No
Catsper1 Exon1	1	Pass	Pass	JTT+I+G+F	Unresolved	No
Catsper1 Mammals	1	Pass	Pass	JTT+G+F	Unresolved	No
Col1a1	1	Pass	Pass	JTT+G	Unresolved	No
Ph20	1	Pass	Pass	JTT+G+F	Resolved	Yes
Porimin	1	Pass	Pass	JTT+G+F	Unresolved	No
Prkar2a	2	Pass	Pass	JTT+I+G	Unresolved	No
Semg2	1	Pass	Pass	JTT+G+F	Unresolved	No
Sp56	2	Pass	Pass	JTT+I+G	Unresolved	No
Zp2	1	Pass	Pass	JTT+G	Unresolved	No
Zp3	1	Pass	Pass	JTT+G+F	Unresolved	No

Genes with significant signal are given in the Likelihood mapping, or, L.M. Category column, see text for explanation of the category 1 and 2 in this column. Results of the amino acid composition and nucleotide base composition bias tests, are shown in the A.A. Comp Bias and Base Comp Bias columns respectively. The phylogenetic trees for each gene are drawn using the substitution model described where G = gamma distributed rates across sites, I = invariable sites, F = frequency of amino acids, JTT = Jones Taylor Thornton model. In the case of LBA analysis, No = no evidence of LBA in the gene analysed, Yes = evidence of LBA in the gene analysed.

were Prkar2a and ZP3 where the presence of polytomies in the gene trees caused the preference of the unresolved nodes over the resolved nodes.

1. Tests of Data Quality and Bias

(i) Test for amino acid and base composition biases

We tested all multiple sequence alignments (MSAs) for evidence of significant levels of amino acid composition bias and base composition bias in each lineage using the TreePuzzle software [15]. We found that all alignments passed the significance test with p-values < 0.05, see Table 1 for summary. For full set of amino acid and base composition bias test results, see Additional Files 1 and 2 respectively. In summary the discordance between each of the gene trees and the canonical species phylogeny is not a result of amino acid or base composition biases providing evidence of false relationships.

Table 2 Summary of SH tests for complete gene datasets

Gene	SH - gene	SH - ideal	Best-fit Tree
Adam2	1.0000	0.1200	NS
Catsper1 Exon1	1.0000	0.1460	NS
Catsper1 mammals	0.5020	1.0000	NS
Col1a1	1.0000	0.2650	NS
Ph20	1.0000	0.3220	NS
Porimin	0.4040	1.0000	NS
Prkar2a	1.0000	0.0490	gene
Semg2	1.0000	0.1010	NS
Sp56	1.0000	0.2380	NS
Zp2	0.1620	1.0000	NS
Zp3	1.0000	0.0050	gene

For each gene, the likelihood of estimated Bayesian phylogeny (gene) and corresponding ideal species tree (ideal) to fit the dataset were determined with the SH test at a 5% significance level. Values equal to 1.0000 represent the tree with the lowest log likelihood, values less than 0.05 refer to those cases where there is a significant difference between the two topologies, and the gene tree is a significantly better fit to the data. NS = No Statistical significance between gene and species tree, in these cases the species tree was used.

(ii) Test for phylogenetic signal

We performed the likelihood mapping procedure implemented in the TreePuzzle software [15,16] to determine the level of phylogenetic signal/conflict present in each alignment, for more detail see the *Methods* section. Our initial dataset consisted of 27 genes, we used this filtering step to reduce our dataset to contain only those genes with phylogenetic signal. We categorized the results from the likelihood mapping analysis into 3 main categories of signal: category 1 had strong phylogenetic signal (see Figure 3a), category 2 had medium level of phylogenetic signal (see Figure 3b) and category 3 had low/no levels of phylogenetic signal (see Figure 3c). The results of the test for phylogenetic signal are summarized in Table 1 and in total 9 out of the 27 genes had strong phylogenetic signal (category 1), with an additional 2 genes with moderate levels of phylogenetic signal (category 2). The complete set of results for the likelihood mapping process is given in Additional File 3. The remaining 17 genes failed the test (category 3). The category 3 genes (with low or no levels of phylogenetic signal) were subsequently removed from the analysis, only 10 genes were retained for further analysis.

(iii) Long Branch Attraction (LBA) analysis

We assessed the data for evidence of LBA which would manifest itself in the data by drawing species with a greater number of mutations in the gene of interest together erroneously on the phylogenetic tree. The method applied uses the MSA and the corresponding phylogeny to categorise rates amongst sites, using an

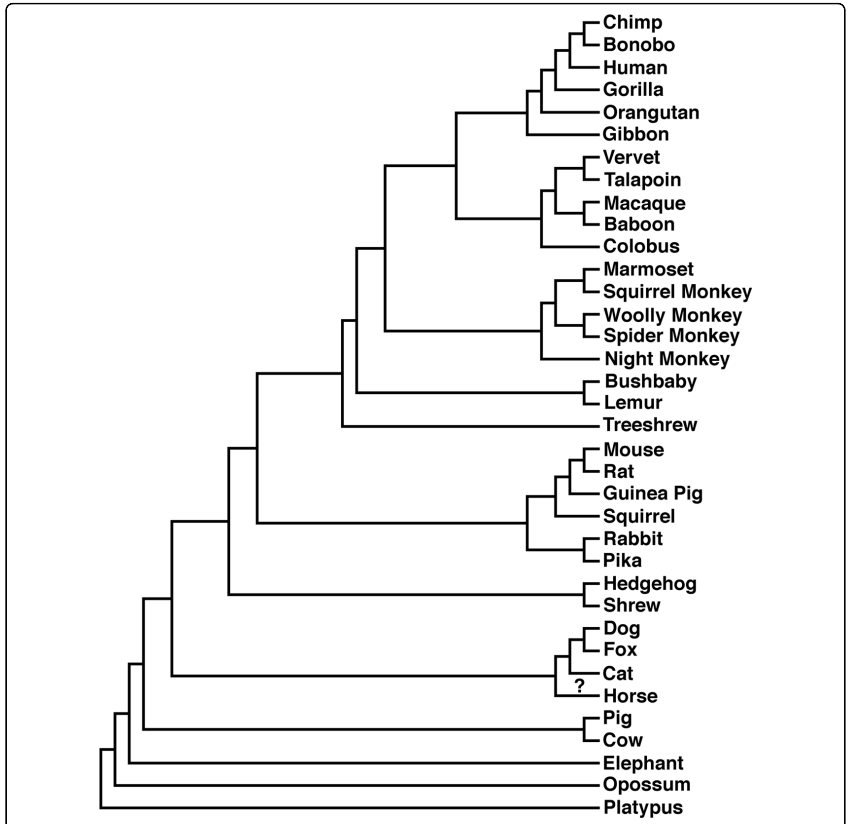


Figure 1 Canonical mammalian species phylogeny. Shown here is a representation of the agreed relationships amongst the *mammalia* for the species used in this analysis. The "?" on the lineage leading to horse indicates controversy over the position of this lineage on the phylogeny.

approach we have previsouly published for mammalian data [17], as described in detail the *Methods* section. In this method of site-stripping we apply the phylogenetic tree (estimated *ab initio* in this software) and the MSA to classify all sites in the alignment into one of eight categories of mutation rate. These are arbitrary categories from 1-8; with 1 being the most highly conserved sites and 8 being the most highly variable. Essentially, these estimates allow us to select only the most conserved sites for phylogeny reconstruction. Sites are sequentially stripped from the alignments based on their rate of evolution and phylogenies are created based on slower evolving sites. These site-stripped phylogenies are then compared to the species tree. Using two independent methods of comparison we determined whether the resultant stripped trees had topologies significantly similar to the species phylogeny. The "root mean squared deviation", or RMSD, method is restricted to binary trees [18], see Additional File 4 for full set of results. Therefore we also employed the SH method of

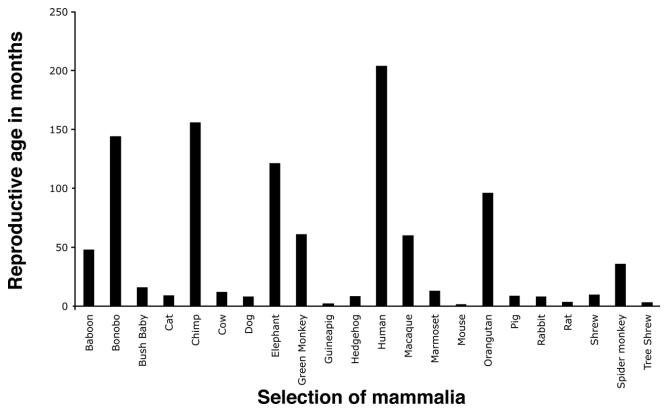


Figure 2 Selection of mammalia used in the analysis and the time to reproductive age in months. Species are shown on the X-axis in alphabetical order. On the Y-axis is the number of months it takes for each species to reach reproductive age.

comparing phylogenies [14], see Additional File 5 for full set of results. For a full description of the RMSD statistic used here [18], see the corresponding section in the *Methods*. Using this approach we could identify the signature of LBA in the Ph20 dataset alone, see Table 1 for summary.

2. Analysis of selective pressures using codon models of evolution

Following analysis of the phylogenies of these reproductive genes, we determined the selective forces at work on these 10 genes (11 datasets). Only those genes passing the data quality tests were analyzed here (*i.e.* 10 genes), see Table 1. In the case of Catsper1, we have analyzed the gene at two different evolutionary distances because it contains high levels of insertion and deletion events. The two datasets for Catsper1 are: exon 1 from the primates only, and, the entire gene from only distant mammalian groups. Hence the number of datasets is 11, and the number of genes tested is 10. The alignments in all cases reached significant levels following randomization tests (z-scores > 1000 in all cases, a z-score of greater than 5 is typically taken as significant).

In those cases where the genes had already been analyzed in previous studies, we expand upon the data in these studies and use more sophisticated models of evolution. ML methods are sensitive to sample size with a minimum of 6 taxa recommended from simulation studies [10]. For a summary of the site-specific and lineage-specific results, see Table 3 and Table 4

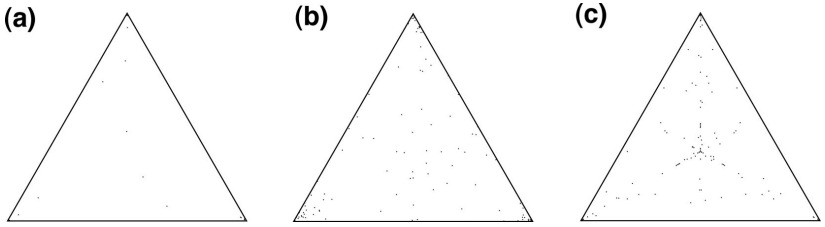


Figure 3 Example of likelihood mapping categories. **(a), category 1** genes with strong phylogenetic signal the example given here is the ZP2 gene, **(b) category 2** genes with intermediate levels of phylogenetic signal, the example given here is the Prkar2a gene, and **(c), category 3** genes with low/no phylogenetic signal, the example given here is the CD9 gene.

respectively. For a summary of all likelihood ratio tests (LRTs) performed in the analyses of these genes see Table A9. In general the lineages tested in the lineage specific analysis for each gene were as follows: modern human; the primate ancestor; modern mouse, and the rodent ancestor, these are indicated in Figure 4(a-k). For certain datasets the species tested varied depending on those species for which high quality sequence data existed for that gene, these are discussed on a gene-by-gene basis below. In summary, for each of the 11 datasets tested, positive selection was detected. In the site-specific test between 7 and 94 sites per gene were identified as positively selected. In the lineage-specific analyses there were up to 2 lineages per gene identified as having evidence of positive selection. Below is a brief description of the results on a gene-by-gene basis, the complete set of all parameters, likelihood values and LRTs are given in Additional File 6.

Col1a1

Possibly the most intriguing result from our entire analysis is that from the Col1a1 protein. According to the microarray study employed here [11], Col1a1 is highly expressed in the uterus tissue. It is also found in most structural tissues including cartilage, bone, tendon, skin and part of the eye (sclera). It is a member of the group 1 collagen proteins involved in the development of the uterine fibroids [19]. There are two propeptide regions to the Col1a1 gene, denoted N- and C-terminal propeptides. According to studies on Col1a1 function, a role has been established for Col1a1 in spermatogenesis [12].

Our site-specific analysis shows 66 sites evolving with an ω value of 4.09, see Table 3. In summary 35/66 of our positively selected sites fall in the N-terminal propeptide region (23-161) and 9/66 positively selected sites fall in the C-terminal propeptide region (1219-1464), this can be seen clearly in Figure 5a. Position 162 in

Table 3 Summary of the results of the site-specific analysis: in each case the most significant model was M8

Gene	n	Parameter estimates	# Positively selected Sites
Adam2	12	p_0 = 0.92632 p = 0.37637 q = 0.60688 p_1 = 0.07368 ω = 3.94326	45>0.50 15>0.95 5>0.99
Catsper1_Exon1 (primates only)	16	p_0 = 0.82736 p = 0.13661 q = 0.03850 p_1 = 0.17264 ω = 3.13071	95>0.50 7>0.95 1>0.99
Catsper1_Mammals (non-primate mammals only)	8	p_0 = 0.83315 p = 0.34233 q = 0.51278 p_1 = 0.16685 ω = 3.26879	124>0.50 30>0.95 8>0.99
Col1a1	10	p0 = 0.98023 p = 0.04796 q = 0.32286 p1 = 0.01977 ω = 4.09285	66>0.50 21>0.95 8>0.99
Ph20	11	p_0 = 0.87658 p = 0.56141 q = 0.83349 p_1 = 0.12342 ω = 2.20500	39>0.50 3>0.95 0>0.99
Porimin	10	p_0 = 0.85067 p = 0.41864 q = 0.32952 p_1 = 0.14933 ω = 12.21841	30>0.50 13>0.95 5>0.99
Prkar2a	17	p_0 = 0.95102 p = 0.16339 q = 0.98823 p_1 = 0.04898 ω = 2.60992	19>0.50 4>0.95 0>0.99
Semg2	12	p_0 = 0.97236 p = 0.01163 q = 0.00500 p_1 = 0.02764 ω = 12.26405	41>0.50 5>0.95 2>0.99
Sp56	14	p_0 = 0.98807 p = 0.16114 q = 1.12262 p_1 = 0.01193 ω = 3.81710	8>0.50 2>0.95 2>0.99
Zp2	18	p_0 = 0.87339 p = 0.63945 q = 0.75356 p_1 = 0.12661 ω = 2.04655	52>0.50 9>0.95 6>0.99
Zp3	13	p_0 = 0.91489 p = 0.30029 q = 0.77328 p_1 = 0.08511 ω = 1.92305	48>0.50 0>0.95 0>0.99

Following LRT analysis M8 was chosen in each case as the most significant model. n refers to the number of taxa in each dataset. The proportion of sites (p), evolving under each corresponding selective pressures (ω) are shown. For example, p_0 refers to the proportion of the protein evolving under the selective pressure value given by ω_0. The parameters p and q describe the beta distribution. The final column gives the number of sites with posterior probability (PP) of 0.50, 0.95 and 0.99 that belong in the positively selected category or sites. The number before the ">" refers to the number of sites with a specific PP value.

Table 4 Summary of lineage-specific positive selection detected.

Species tested as Foreground	Significant LRT	Parameter estimates		
		P	Fwd ω	Bck ω
Adam2				
Macaque	ModelA v M1	9.57%	1.71	0.10/1
Catsper1 Mammals				
Ferungulata	ModelA v M1	4.46%	998.99	0.09/1
Rodents	ModelA v M1	5.45%	999.00	0.084/1
	ModelB v m3Discrtk2	4.47%	999.00	0.12/1.38
Col1a1				
Rodents	ModelA v M1	2.17%	72.73	0.013/1
	ModelB v m3Discrtk2	1.93%	72.77	0.02/1.35
PH-20				
Guinea Pig	ModelA v M1	6.3%	11.48	0.13/1
	ModelB v m3Discrtk2	6.14%	12.57	0.14/1.10
Prkar2a				
Macaque	ModelA v M1	2.37%	999.00	0.04/1
	ModelB v m3Discrtk2	2.53	999.00	0.04/1.22
Sp56				
Human	ModelB v m3Discrtk2	100%	62.40015	0.02/0.55
Glires	ModelB v m3Discrtk2	2.56%	1.03	0.02/0.55

Summary table of significant results for lineages specific analyses following LRT analyses. Lineages tested as foreground (Fwd) are shown in the first column. Only those lineages with significant LRT values for Model B or Model A and ω >1 are shown here. Parameter estimates are given for the LRT values highlighted in bold. P is the proportion of sites under selection the corresponding selective pressure as measured by ω. Fwd ω and Bck ω scores for the foreground species and background species respectively are given in the final column.

Col1a1 is cleaved and modified by an endopeptidase, position 162 is also modified by pyrrolidone carboxcylic acid (Swiss-Prot PO2452). A positively selected site at position 163 is neighboring this multifunctional site, suggesting that there has been an evolutionary effort to improve cleavage and/or modification in this protein.

Variations in Col1a1 are linked with Osteogenesis Imperfecta (OI), an autosomal dominant disease, resulting in an inability to make the correct collagen protein. There are a spectrum of OI conditions, the most severe is OI type 2 (OI-II) leading to death in the perinatal period. A recent extensive study of the Single Nucleotide Polymorphisms (SNPs) associated with OI has revealed a number of substitutions of glycine residues within the triple helical domains of the Col1a1 protein [20]. The total number of disease implicated sites in the Swiss-Prot entry P02452 for Col1a1 is 99: 4 of these are OI non-specific, 4 are OI-I, 59 are OI_II, 14 are OI-IV and 15 are SNPs (2 are associated with another disease). One third of the mutations that result in substitutions for glycine in Col1a1 are lethal whereas those between the start codon and position 200 are non-lethal. Only 1 of the sites we have identified as positively selected is in the non-lethal domain from position 1-200, this is site 195. This positively selected site is neighboring the SNP position 197 that causes a mild OI phenotype. In Table 5 we show a list of 11 positively selected sites that fall in close proximity to sites associated with disease and are located between 280 and 1456, spanning the important triple helix region. These positions are all within 1 to 5 amino acid residues of known disease variants, 8 of these disease variants are the severe/lethal OI-II disease form. Two exclusively lethal regions, helix positions 691-823 and 910-964 aligned with major binding regions [20] and we find a positively selected site in this region. Following a randomization test for the positively selected sites and disease implicated sites (as denoted by Swiss-Prot entry PO2452), we have found that the pattern we observe, i.e. finding positively selected sites in close proximity to disease implicated sites is significant in 3 out of the 11 cases examined here (at $P < 0.05$).

Lineage-specific analysis shows evidence for positive selection in this protein in the rodent ancestor. In total, 2.2% of the sites in the rodent ancestor have $\omega = 72.73$, while the rest of the species are evolving under purifying selection, $\omega = 0.013$. For a summary of site and lineage specific results for Col1a1, see Table 3 and 4. For complete set of results see Additional File 6(d).

Prkar2a (interacts with SEMG2)

Prkar2a is a cAMP dependent protein kinase that is attached to the sperm flagella via regulatory subunit (RII) [21]. Protein tyrosine phosphorylation has been linked with successful fertilization due to hyper-activated sperm motility [22]. This increase in phosphorylation is part of a cAMP dependent pathway that activates protein kinase A [22].

The PRKA families were previously tested for positive selection using 3 to 4 taxa and site-specific model M8 with no significant results for positive selection reported. With our 17 taxa dataset, we were able to detect that 4.7% of sites were evolving at a rate of $\omega = 2.60$, see Table 3 for summary of details.

Positively selected sites detected in the site-specific analysis of Prkar2a were compared to the human Swiss-Prot sequence (P13861). In total 18 sites were predicted to be positively selected, 17 of these sites occur in the region of the protein associated with dimerization and phosphorylation (2-138), see Figure 5(c). In the Swiss-Prot entry there are a number of residues listed as being modified by phosphoserine. These are positions 58, 78, 80, 99 and phosphothreonine at position 54. The sites estimated to be positively selected from our analysis are: 58, 59, 61, 62, 63, 64, 65, 68, 70, 74, 75, these sites are at or in close proximity to these modified residues.

The regulatory subunit alpha 2 of Prkar2a has been shown *in vitro* to interact with Semg2. The phosphorylation of Semg2 may lead to its activation into forming a gel matrix in the female reproductive tract. From our analysis it is shown that while Semg2 has positively selected sites dispersed throughout its sequence, whereas the positively selected sites for Prkar2a are localized to the N-terminus region, and the remainder of the gene is under strong purifying selection. Literature has so far not specified an exact phosphorylation site for Semg2, which prevents us from commenting further on its interactions with Prkar2a.

Lineage-specific analysis shows that Prkar2a in the macaque has undergone a greater selective pressure to change when compared with other *mammalia* in the dataset, with 2.53% of sites evolving at $\omega = 1.22$, see Table 4 for summary of results. For complete set of results for Prkar2a, see Additional File 6(g).

Ph20 (interacts with ZP2 and ZP3)

Ph20 is expressed in the testis and found in the acrosome of the sperm. It is also codes for a receptor that is involved in the sperm to zona pellucida (ZP) adhesion [23].

Previous analysis conducted on this protein involved 6 taxa [24]. Here we have increased the number of taxa to 11. We have omitted the carnivores from our analysis of Ph20 as the sequences were spurious. We found evidence for LBA in the Ph20 dataset. By removing fast evolving sites a fully resolved gene phylogeny is obtained. This gene tree now is in agreement with the ideal species phylogeny ([13].

Lineage-specific analysis shows that guinea pig is under positive selection, with 6.1% of sites with $\omega = 12.57$ while all other species in the background are evolving at $\omega = 0.14$ or neutrally, see Table 4. The 39 positively selected sites were then compared to the human Swiss-Prot sequence (P38567), see Figure 5(b) for results. Catalytically important resides 146, 148, 211 284 and 287 when mutated result in a reduction in, or loss of, activity [25]. It has been shown experimentally that mutations in the region of this active site significantly reduce or completely block the function of this protein [25]. Our results show that 3 of the positively selected sites, 155, 272, 273, are in close proximity to these regions. Another 5 positively selected sites: 83, 155, 252, 353 and 391 are close to glycosylation sites, see Figure 5 (b). These sites when modified are known to change the structure and function of the Ph20 protein. For complete set of results for Ph20 see Additional File 6(e). These results are of significance as the Ph20 protein changes position in the sperm during the different stages of the fertilization process. In guinea pig Ph20 protein is known to migrate from the post acrosomal membrane to the inner acrosomal membrane [26]. Thus finding these positively selected sites in close proximity to these glycosylation sites in guinea pig suggests that these sites have been selected to modify the Ph20 structure more effectively thus increasing the chance of capacitation.

SP56 (interacts with ZP2 and ZP3)

The binding of sperm to the zona pellucida (ZP) is crucial for gamete formation to take place. The exact mechanisms of this process are still to be uncovered therefore any predictions on important residues will greatly improve knowledge by directing mutational studies. SP56 has been shown through photoaffinity cross-linking experiments to have a specific binding affinity for ZP3 [27]. Therefore it is believed to play an important role in the binding of sperm to the ZP matrix. Experiments have shown that during capacitation SP56 is released from the acrosomal matrix and becomes situated in the sperm head membrane, enabling it to act as a ZP3 binding protein [28].

Here we have found 8 positions in the SP56 protein that are under positive selection ($\omega = 3.82$) following site-specific analysis. These sites were compared to the human SP56 entry in Swiss-Prot (Q13228) to determine possible links to function. One of these 8 positively selected sites is position 122, regarded as a SNP number (rs35396382) in dbSNP database [29]. Although further experimental work needs to be conducted to decipher the clinical association of this position, it is extremely interesting that our most significant positively selected site also displays variation in the population, especially given the overall high level of conservation in this gene. For summary of results see Tables 3 and 4, and for full set of results for this gene see Additional File 6(i).

ZP2

Zona pellucida (ZP) proteins form the complex glycoprotein coat that surrounds the oocyte [30]. These ZP

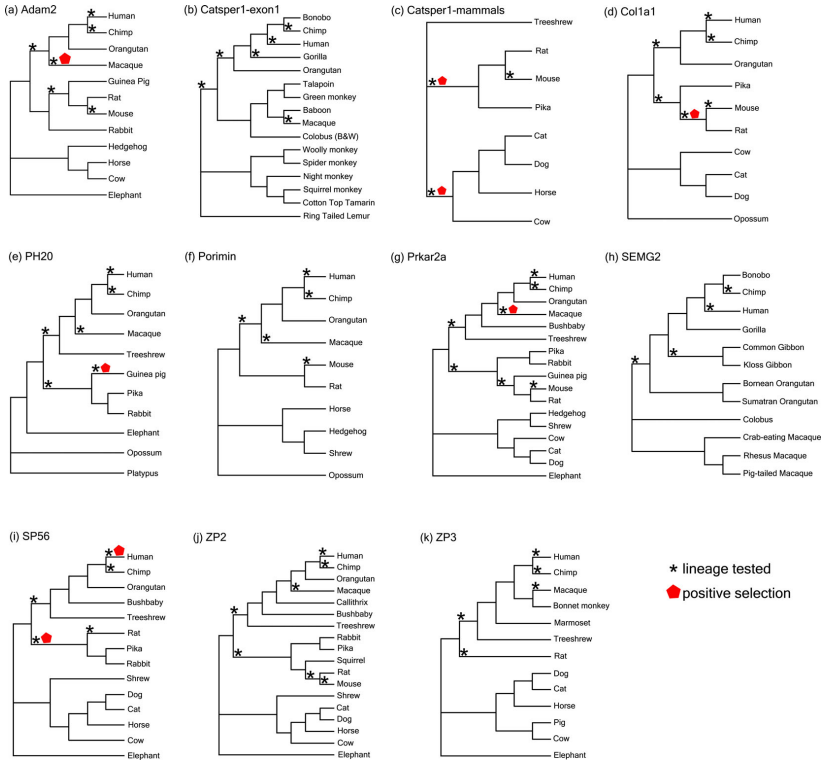


Figure 4 Results of lineage specific positive selection analysis on 11 datasets. The phylogeny used for each gene is a reduced version of the species phylogeny. The lineages labeled as foreground in the analysis are denoted in the diagram with asterix symbols. These are the results following LRT analysis. Those lineages where positive selection was determined are represented by red pentagons.

proteins have been shown to be under strong pressure to change, and results have been published on both site and lineage analyses [31]. Here we have expanded the analysis of ZP2 to include 18 taxa (maximum previously tested = 8 [31]). We have also applied more complex models of evolution and have sampled deeper branches on the phylogeny including a representative of the *Afrotheria* - elephant.

In this case, the results of our larger dataset and more complex models show that the values of ω determined here vary slightly when compared to previous analyses [31]. This previous test showed 4.7% of sites to have ω = 2.5, increasing the size of the dataset in this study results in 52 sites in ZP2 that have an ω value of 2.05. See Additional File 6(j) for complete results.

Positively selected sites were compared to the human Swiss-Prot entry for ZP2 (Q05996) to identify possible function for these sites, see Figure 5(d). ZP2 contains 7 carbohydrate chains situated between sites 87-462, these are important for the sperm to bind to the ZP of the egg coat [32]. Of the 46 sites identified to be under positive selection, 23 fall between positions 66-257, this region contains 5 of the binding domains of the carbohydrate chains. The clustering can be seen more clearly

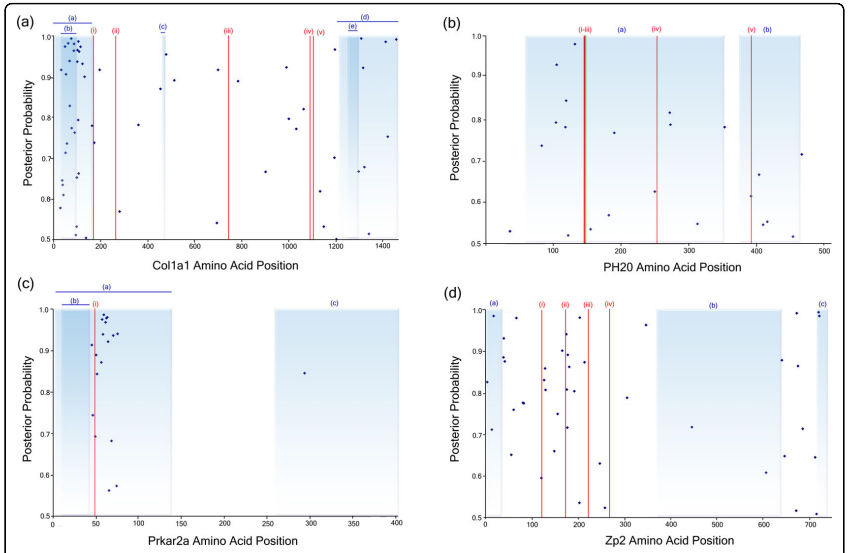


Figure 5 Results of positive selection analysis for 4 genes. Each of the four graphs represents the CDS of a gene from position 1 to the stop codon (X-axis). The Y-axis is the posterior probability of each of the sites belonging to the positively selected category. The dark blue data points are sites estimated to be under positive selection. Alternative pale blue and white regions depict alternative domains in the protein, this data is taken from Swiss-Prot. The vertical red bars in each case represent functionally important sites, these are specific to each gene as follows: **(a) Col1a1**, (i) Cleavage site by procollagen N-endopeptidase, (ii) O-linked Gylcosylation site, (iii) Cell attachment site, (iv) Cell Attachment site, and (v) O-linked Glycosylation site. **(b) PH20**, (i) active site, proton donor, (ii) and (iii) are positions when mutated result in loss of activity, and (iv) N-linked Glycosylation site, **(c) Prkar2a**, (i) and is a Phosphothreoinine modified residue, **(d) ZP2**, (i), (iii) and (iv) are N-linked Glycosylation sites, and (ii) is the cleavage site.

in Figure 5(d). Another cluster of positively selected sites (10 sites in total) occurs in the propeptide region (641-745). It has been suggested that upon the cleavage of the propeptide region, the mature ZP2 protein plays a role in the prevention of polyspermy [33].

ZP3

Analysis of site-specific evolution in ZP3 identified 48 positively selected sites. Of specific interest are positively selected positions 329, 330, 332, 336, 338, 339, as these sites were in close proximity to identified sperm binding sites (329-334) [34], see Table 3. The furin cleavage site is identified at position (350-353), and the propeptide domain at position (351-424). When cleavage takes place the ZP3 undergoes a conformational change that inhibits any further sperm binding to the coat thus preventing polyspermy [35]. Of the 48 positively selected sites identified, 10 fall within the propeptide domain, with an additional 12 occurring close to the vicinity of the furin cleavage and sperm binding sites, thus suggesting that there is a pressure to improve binding and prevent polyspermy. For complete set of results for ZP3, see Additional File 6(k).

Adam2 (Fertilin β)

Adam2 is a cell adhesion molecule that plays a fundamental role in the final binding of sperm to the oocyte membrane [36]. Indirect interactions have been shown with female proteins CD9 [37]. (We have not continued further analysis on CD9, as it failed the likelihood mapping test).

Previous results have been published reporting positive selection using site-specific analysis on 6 taxa [24]. Here we have included 12 taxa for Adam2 and we have investigated the possible functional implications of positively selected sites found. In the site-specific analysis we find 7.3% of sites with ω = 3.94, this corresponds to 45 sites in total, see Table 3. Comparison of these positions to human Swiss-Prot Adam2 sequence (Q99965), we determine that 39/45 positively selected sites are

Table 5 Summary of the positively selected sites in the col1a1 gene, their clinical relevance, and, the probability of being located within distance "d" from the nearest disease-implicated site.

Positively selected sites	Posterior Probability	Human Variant: SNP position	Distance (*d*)	Probability of being *d* from nearest disease-implicated site	Genetic code distances between observed character states	Clinical Association
195	0.926	197	2	0.04	A-N = 2	G → C mild phenotype
280	0.588	275	5	0.26	A-S = 1; S-T = 1; T-A = 1	G → D OI-II
478	0.959	476	2	0.128	A-S = 1; S-T = 1; T-A = 1	G → R OI-II
784	0.968	776	8	0.396	A-S = 1; S-T = 1; T-A = 1	G → S OI-II
1032	0.535	1025	7	0.364	A-P = 1	G → R OI-II
1063	0.826	1061	2	0.128	N-S = 1	G → D OI-II
		1061	2	0.032	N-S = 1	G → S OI-IV
1149	0.623	1151	2	0.032	A-S = 1	G → S OI-III
		1151	2	0.128	A-S = 1	G → V OI-II
1194	0.675	1195	1	0.076	A-G = 1; G-S = 1; S-A = 1	G → C OI-II mild form
1196	0.972				A-F = 2; F-Y = 1; Y-A = 2	
1316	0.928	1312	4	0.24	K-N = 1; N-P = 2; P-K = 2	W → C OI-II
1456	0.997	1460	4	0.1	C-F = 1; C-L = 2; C-M = 2; F-L = 1; F-M = 2; L-M = 1	P → H dbSNP: rs17853657

The sites under positive selection in the col1a1 protein and their associated posterior probabilities (PP) are shown. The third column shows variant positions (SNPs) as determined using Swiss-Prot human (PO2452) sequence. The fourth and fifth columns show the residue distance "d" of the positively selected site from its nearest genetic variant, and the probability of being located "d" residues from any disease implicated site by random chance alone. The sixth column uses single-letter amino acid symbols to show the genetic code distances between all observed character states at each positively selected site. "Clinical Association" show the replacement substitution at the human variant position and its clinical association with that human variant. OI = Osteolysis imperfecta, OI-I to -IV. The final entry for dbSNP is database entry number rs17853657 and as yet has not been associated with OI although it is in the same domain as the other disease-causing SNPs.

situated in the C-terminus region. On closer investigation of these sites we find that 12/45 positively selected sites occur in the disintegrin domain (position 384-473). The disintegrin domain has been shown to be involved in the binding of Adam2 to the oocyte [38]. A cysteine-rich domain occurs between (477-606), 16/45 positively selected sites fall in this region. It has been suggested for Adam12, (another member of the Adam family of proteins), that the cysteine-rich domain plays a role in mediating the cellular interactions via syndecans and integrin [39], a similar role for this domain in Adam2 can be postulated. Overall the results for Adam2 suggest a selective pressure for increased binding of Adam2 to the oocyte regardless of species of origin. For a complete set of results and LRTs for Adam2, see Additional File 6 (a).

Catsper1

Catsper1 is involved in regulating the calcium cation channel in sperm flagella, the result of which is movement of sperm [40]. Previous studies on Catsper1 exon 1 have been performed [41]. We intended to expand our analysis to span all exons and expand the data set to include a variety of *mammalia*. However, the exon 1 of non-primate *mammalia* is so highly variable that an accurate alignment cannot be constructed. The remaining exons were highly conserved across all species. We therefore split our catsper1 dataset into two sections each of which produced a good quality alignment for analysis, (1) exon1 of catsper1 for the primates, and (2) entire catsper1 gene for non-primate *mammalia*.

(a) Catsper1 Exon 1 primatesSite-specific analysis of this protein identified 17% of the protein under positive selection with ω = 3.13. Previous analysis of this exon showed positive selection on indel substitutions in this gene [41]. The positively selected sites are situated throughout exon1, little is known about the functional significance of these sites. However, it is known that exon 1 has a significant role to play in altering the rate

of calcium ion channel inactivation. Different lengths in the N-terminus result in different rates of channel inactivation, where a long terminus results in a longer time to activation than the shorter terminus. This is described most effectively by the ball and chain mechanism described in [41]. See Additional File 6(b) for complete results. These results show the importance of this protein, and specifically the first exon, for reproductive success.

(b) Catsper1 entire gene non-primate mammalsOur site-specific analysis identified 16.7% of the sites under positive selection with an ω = 3.27, see Table 3. These sites all cluster in exon 1. While the rodent ancestor appears to be under positive selection with 4.47% of its sites evolving at ω = 999, see Additional File 6(c) for complete set of results. A previous study of 9 rodent species, including *Mus musculus* individuals from 4 different populations, has shown that within the rodent order there has been a continued pressure to evolve, with positive selection for indel substitutions in exon1 of the Catsper1 gene [43].

Semg2

A member of the family of semenogelin genes, Semg2 is involved in the formation of a postcopulatory plug [44]. Previously, positive selection has been reported for both site-specific and lineage-specific analysis for Semg2 [9,45]. We have expanded the data set from previous analyses to incorporate more species.

In our site-specific analysis, we found that 2.7% of our sites had an ω value of 12.26, see Table 3.

We have performed a novel functional analysis of these positively selected sites by comparing them to the human Semg2 sequence (Q02383) in the Swiss-Prot database. This is a step not previously taken by other studies of Semg2. A striking pattern emerged - all known domains of this protein have several positively selected sites. There is a probable glycosylation site at position 272, which is located close to a large stretch of positively selected sites (positions 262 to 289). It is so far unknown how significant this glycosylation site is in Semg2 and whether it plays a role in modifying the protein to form a copulatory plug. However, the results indicate that this protein, and in particular the region around the glycosylation site, has been under significant pressure to change.

A complete set of results for Semg2 is given in Additional File 6(h). The lineage-specific results are not described here in detail as lineage analyses have been carried out previously on the primate Semg2 gene [9,45]. It has been shown recently that the rate of evolution for this protein varies depending on the level of sperm competition [9]. Our results are in agreement with this finding, thus further verifying our approach.

Porimin

Two isoforms of this protein have been identified; we have focused on isoform 1 in the *mammalia*, as isoform 2 contains an additional human specific region between residues 34-52. To date the exact mechanisms of this transmembrane receptor are unknown. This protein is not well characterized biochemically and its function cannot be verified as reproduction related, therefore we only discuss the results briefly below.

On site-specific analysis of this protein we determined that 30 of the sites are under positive selection (ω = 12.22), see Table 3. From analysis of the sites on the Swiss-Prot entry for human Porimin (Q8N131), we could determine that two positively selected sites (146 and 147), were found in a highly conserved region and fall in close proximity to the N-linked glycosylation site. For complete set of results for Porimin, see Additional File 6(f).

Conclusion

Testing for phylogenetic signal and biases, such as amino acid composition bias and LBA, indicated that there was adequate phylogenetic signal for 10 of the genes and in general no evidence of systematic biases. On testing for LBA, Ph20 was the only protein in this dataset that displayed the typical signature of this bias with gene and species tree agreement being maximized with the removal of the fastest evolving categories. This would suggest that while germ line generation times vary greatly in the dataset, the effect of the resultant LBA does not impact on the sequence data to any great extent (1/11datasets).

Selective pressures for the reproductive proteins studied here are heterogeneous. All proteins exhibited regions of strong conservation proving the importance of maintaining structural stability and overall function in these proteins. All but 1 protein (Adam2) exhibited evidence of positive selection in specific lineages, and all proteins without exception exhibited positive selection in regions of catalytic/functional importance. For SP56 and Col1a1 the site-specific results are entirely novel. The lineage-specific results described here for Prkar2a and Catsper1 exon 1 in primates, are also novel. We have shown that, in the case of Catsper1, there is a fundamental protein functional shift between new world monkeys and old world monkeys. The Dn/Ds measurement applied here assumes that neutral substitution rate is akin to Ds, therefore no selection on silent sites. There have been many publications of late to the contrary therefore we are mindful of examining the rate of silent substitution in all our analyses [46,47].

For the reproductive genes in our dataset, we show that lineages evolve at unique rates and at functionally

crucial sites, specifically those involved in phosphorylation. We have also shown that a number of these proteins (Col1a1 and Catsper1) show positive selection for example in the ancestral rodent lineage and evidence of purifying selection in the subsequent divergent species.

Overall our analyses of these reproductive proteins show how important it is to carefully examine data for systematic biases prior to testing for lineage and/or site specific positive selection. We have also demonstrated the importance of including large numbers of taxa/lineages in these analyses. This finding was highlighted in our analysis of Prkar2a where previous analysis of this protein had included only 4 taxa and therefore reported a negative result. We do not observe any large-scale effect of germ line generation time in our dataset, with only 1 protein (Ph20) with evidence of long branch attraction. The results of Col1a1 indicate that the positively selected sites may have been of such importance for this protein that neighboring mutated sites may have been maintained in the population despite their propensity for causing disease. The location of positively selected sites determined using this approach and in regions of functional importance in the proteins in this dataset, provides us with further evidence of the link between functional shift and positive selection.

Methods

The data analyzed in this study consist of homologous reproductive genes from a variety of mammalian genomes. Genes were identified as being reproduction related from literature searches, analysis of protein interaction networks (iHOP) [48] and expression (microarray) data [11]. The microarray expression data used is from normal human tissues. We have also included a more in-depth analysis of previously identified cases of positive selection in reproductive proteins. A list of all data used in this study are available in Additional File 7, the total number of genes analyzed was 10. Homologs of all 10 reproduction related genes were identified in mammalian genomes that span the entire phylogeny of mammals, see Figure 1. For each of the reproduction related genes, the alignment of homologs contained between 10 and 18 species, and the alignment length varied between 351 and 4374 base pairs.

Sequence Data

Protein coding sequences for the reproductive proteins were retrieved by the combination of two methods; Ensembl and Blast searches. Orthologous coding sequences from all available completed mammalian genomes were retrieved from the Ensembl database [49]. These orthologs had been identified previously by performing a genome-wide reciprocal WUBlastp +SmithWaterman search of each gene across all completed genomes. To include those *mammalia* that were not present in Ensembl a BlastP search was conducted on all the human amino acid sequences from each gene against the Swiss-Prot database.

Mammalian Species

Primates: Human *(Homo sapiens)*, Chimp (*Pan troglodytes)*, Bonobo *(Pan paniscus)*, Bornean Orangutan *(Pongo pygmaeus)*, Sumatran Orangutan (*Pongo abelii*), *Gorilla (Gorilla gorilla)*, Rhesus Macaque *(Macaca mulatta)*, Crab eating *Macaque* (*Macaca fascicularis*), Pigtailed Macaque (*Macaca nemestrina)*, Bonnet monkey *(Macaca radiata)*, Baboon (*Papio hamadryas*), Mantled Guereza *(Colobus guereza)*, Vervet Monkey (*Cercopithecus aethiops*), Angolan Talapoin (*Miopithecus talapoin*), Squirrel Monkey (*Saimiri sciureus*), Cotton top tamarin (*Saguinus oedipus*), Common Marmoset (*Callithrix jacchus)*, Marmoset/Callithrix *(Callithrix-jacchus)*, Spider Monkey *(Ateles geoffroyi*), Bushbaby *(Otolemur garnettii*), Common woolly monkey (*Lagothrix lagotricha*), Ringtailed lemurs (*Lemur catta)*, Kloss Gibbon *(Hylobates klossii)*, Common/Lar Gibbon *(Hylobates lar)*, Night/owl Monkey *(Aotus trivirgatus boliviensis)*. Scandentia: Treeshrew *(Tupaia belangeri)*. Rodents: Mouse *(Mus musculus*), Rat *(Rattus norvegicus)*, Guinea pig *(Cavia porcellus)*, Ground Squirrel/Squirrel *(Spermophilus tridecemlineatus)*. Lagomorpha: Rabbit (*Oryctolagus cuniculus*), Pika (*Ochotona princes*). Eulipotyphila: Hedgehog *(Erinaceus europaeus)*, Shrew (*Sorex araneus*). Carnivores: Cat *(Felis catus)*, Dog *(Canis familiaris)*. Artiodactyla: Cow *(Bos taurus*), Pig *(Sus scrofa)*. Perisodactyla: Horse *(Equus caballus)*. Proboscidea: Elephant *(Loxodonta africana)*. Monotremata: Platypus *(Ornithorhynchus anatinus)*. Didelphimorphia: Opossum *(Monodelphis domestica)*.

Multiple Sequence Alignment (MSA)

All coding sequences were translated into their corresponding amino acid sequences using in-house translation software. Gene family alignments were generated at protein level using ClustalX 1.83.1 using default parameter settings [50]. The corresponding nucleotide gene family datasets were aligning based on their protein alignments using in-house software. Each gene family alignment was manually edited using Se-Al [51] to remove any ambiguous regions.

Nucleotide composition bias, amino acid composition bias and likelihood mapping tests

TreePuzzle 5.2 [15] performs a chi-square test that compares the amino acid composition of each sequence to the frequency distribution assumed in the General Time Reversible (GTR) and Jones Taylor Thornton (JTT) models [52]. Ideally no species should fail this test, however, where two species fail and are thus drawn together on a tree, these sequences are excluded. Using the likelihood mapping method, each tree is disassembled into its constituent quartets and the support for each possible quartet is assessed. If the data contains phylogenetic signal then the likelihood of all three possible relationships for that quartet will be equally likely, these are represented by the three tips of the triangle, and the majority of the signal will be in these tip regions. Otherwise, the vertices and central region will be most heavily populated by supporting quartets.

Phylogeny Reconstruction

Phylogenetic trees were constructed using MrBayes v3.2.1 [53] and the amino acid sequences. Amino acid sequences were used in order to vitiate the effects of base and codon compositional biases. The substitution model was selected following model testing using Modelgenerator version 85 [54]. The selected model was JTT, the GTR rate model was implemented and the first 20000 trees for each gene were discarded as "burnin". A majority rule consensus tree from the remaining trees sampled was constructed for each gene. The parameter settings for each gene phylogeny are summarized in Additional File 8.

Site-stripping for significance

To test for long branch attraction (LBA) we applied the slow-fast approach of Brinkman and Phillipe [55]. We implemented the rate categorisation in a maximum likelihood framework in TreePuzzle 5.2 [15]. This software takes the alignment as input and generates *ab initio* phylogenetic trees. It then calculates the rate of mutation for each site in the alignment. The software specifies 8 arbitrary categories of site: each one of these categories contains some portion of the alignment. In this manuscript 8 is the most rapidly evolving (for example every lineage has a different character state for that character), and category 1 is the most slowly evolving (for example each lineage has the same/identical character state for that character). Sites are then progressively removed from the protein MSA according to their evolutionary rate, and at each stage a new phylogenetic tree is constructed based on this slightly reduced dataset. The difference between the new topology created on a reduced alignment and the original topology reconstructed based on the entire alignment are then compared in a statistical framework to determine which fits the data best (SH Test 2, see below) or which is most similar to the species phylogeny (RMSD Test 1, see below). At each stage we employ MrBayes [56] to perform the phylogenetic reconstruction using the aforementioned settings.

Tests of the difference between two trees

Test 1: Nodal distance calculation

TOPD/FMTS v 3.3 [18] calculates the distance between the site-stripped trees and the 'ideal' tree. The 'ideal' tree used for each gene was a pruned version of the canonical species tree as seen in Figure 1. A distance matrix is derived by counting the number of nodes that separate each of the taxa in a tree. A distance matrix is calculated for each site-stripped tree as compared to the ideal species tree. The nodal distance score is obtained by calculating the RMSD of the matrices. If both trees are identical the RMSD value would be 0, indicating no distance between them. This figure increases the more distance there is between the two trees.

Test 2: Shimodaira-Hasegawa (SH) statistical test of two trees

For each gene MSA, complete and site-stripped, a comparison of the likelihood of the estimated Bayesian phylogeny for that alignment with the likelihood of its corresponding 'ideal' species tree was carried out using the SH test [14] implemented in TreePuzzle 5.2 [15] to determine which tree was significantly the best-fit tree for the alignment.

Selective Pressure Analysis

PAML 4.3 [57,58] uses a ML method of calculating ω for site-specific and lineage-site specific changes. Codeml, part of the PAML 4.3 package [57,58], applies a series of models to our data, with each model differing from the previous with the addition of more complex parameters. The simplest model is M0, and it calculates an ω value over the entire alignment. This model assumes that all sites and all lineages are evolving at the same rate. Model M3 is an extension of M0 and allows all ω values to vary freely. There are two variations of the M3 model, m3(k = 2) discrete which allows two variable classes of sites and m3(k = 3) which allows three classes of site. M1 is a neutral model that allows two parameter estimates for proportion of sites where ω = 0 or ω = 1. M2 is the selection model, it allows three parameters where ω = 0 or ω = 1 or ω is estimated and free to be greater than 1. M7, is the beta model, it allows ten different site classes for ω between 0 and 1. M7 is compared against the more parameter rich M8 (beta &omega >1). M8 allows 10 different site classes but contains an additional parameter whereby the 11th ω is free to vary between 0 and >1. M8a(beta &omega = 1) is null hypothesis of model 8. Model A & Model B are models that allow

testing of ω variation in lineage-site analyses. Model A is an extension of M1 and Model B is a more parameter rich extension of m3(k = 2). We have also implemented model A null which is denoted as modelA1 elsewhere. Model A null is compared to model A in an LRT as per Additional File 9. Only statistically significant models for the data are taken into account. Statistically significant results were decided by calculating the difference in log likelihood or, lnL, scores between models and their more parameter rich extensions in a likelihood ratio test (LRT) as described previously in [17,58]. If the likelihood score was exceeded the critical χ^2 values, then the result was significant. See Additional File 9 for full set of LRTs performed.

In silico analysis of positively selected sites

Sites under positive selection (ω > 1) were estimated using the empirical Bayes methods in the site-specific and lineage specific analysis performed. The methods used were naúve empirical Bayes (NEB) and Bayes empirical Bayes (BEB) [58]. Swiss-Prot is a protein sequence database that provides description of the function of a protein, the domain structures, post-translational modifications and variants. Significant sites, verified through close examination of the MSAs and codeml output using alignment visualisation software Se-AL [51], were compared with unaligned human amino acid sequence taken from Swiss-Prot. These sites were examined to see whether or not they lay in catalytically important regions of the protein.

Additional file 1: Additional Table 1 - Results of amino acid composition bias per gene. Results of the amino acid composition bias test and shown here on a per gene basis. We would expect that if two species have similarly and significantly (P < 0.05) biased amino acid composition that they would be drawn together on the phylogeny. Those with P < 0.05 scores are highlighted but are dispersed throughout different genes. The frequency distribution assumed in the maximum likelihood model calculated by Tree-Puzzle (5% chi-square p-values) was used. N/A = species not represented in the gene dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S1.DOC]

Additional file 2: Additional Table 2 - Results of base composition bias per gene. Results of the base composition bias test and shown here on a per gene basis. We would expect that if two species have similarly and significantly (P < 0.05) biased base composition that they would be drawn together on the phylogeny. Those with P < 0.05 scores are highlighted but are dispersed throughout different genes. The frequency distribution assumed in the maximum likelihood model calculated by Tree-Puzzle (5% chi-square p-values) was used. N/A = species not represented in the gene dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S2.DOC]

Additional file 3: Additional Table 3 - Results of likelihood mapping test for phylogenetic support and conflict estimated for each gene. Results of Likelihood mapping test are shown here on a gene-by-gene basis. This table summarizes the amount of phylogenetic signal and conflict in each alignment. The three possible topologies for each quartet of species are represented by the corners of the triangle, these corners represent strong support for phylogenetic signal. Quartets present on the vertices represent incongruence in the phylogenetic signal. Quartets at the centre of the triangle represents those quartets where all three topologies are equally likely, i.e. phylogenetic signal completely lacking. Each gene is subsequently given a category based on the quality of the data, only categories 1 and 2 were used.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S3.DOC]

Additional file 4: Additional Table 4 - Results of root mean squared deviation (RMSD) analysis for comparing binary trees. This table summarizes the results of comparing the site stripped phylogenies with the ideal species phylogeny. In the first column is the gene name. Each of the subsequent columns represents a category of site variation that is removed (1 is the slowest evolving, 8 the most rapid). The values given for each category removed is the RMSD statistic and represents how similar the resultant site stripped topology is to the canonical species phylogeny. NB - non-binary tree, N/A - not applicable (site category not estimated for alignment).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S4.DOC]

Additional file 5: Additional Table 5 - Results of the SH test for site-stripped gene versus ideal species phylogeny. This table summarizes the results of comparing the site stripped phylogenies with the ideal species phylogeny using the SH test, this is a more statistically robust approach and more suited to multi-furcating topologies such as those in the dataset. Each of the rows represents a category of site variation that is removed. For each site stripped site dataset the resultant gene tree is compared to the species phylogeny. The values given for each category removed denotes whether there is a significant difference between the site stripped tree and the species phylogeny, values of less than 0.05 represent those cases where there is a significant difference between the phylogenies. NS = No Statistical significance between gene and species tree, the species tree was taken in these cases.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S5.DOC]

Additional file 6: Additional Table 6(a-k) - Complete results of Maximum likelihood analysis for selective pressure variation per gene. For each gene analyzed (a-k) the results are shown in full on a gene-by-gene basis (in alphabetical order). The layout of each table is identical for each gene. The corresponding LRTs performed and all scores and values computed are shown below. The models used are given in the left-most column (Model), followed by the number of parameters associated with that model (P). The Log Likelihood or each model is given in the column (L), and the estimates of the parameters for the proportion of sites (p) and the ratio of Dn/Ds (ω) are given. Sites identified by each model as being positively selected are shown in the final column.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S6.DOC]

Additional file 7: Additional Table 7 - Summary of data used in the analysis. Species names, unique identifiers and sequence lengths are given for all data. Summary of data used in the analysis. Species names, unique identifiers for Ensembl (ENS) or Swiss-Prot and database versions are given. The sequence length per species are given for all genes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S7.DOC]

Additional file 8: Additional Table 8 - Parameters for Phylogeny Reconstruction per gene. The parameters used to reconstruct each gene tree in MrBayes are shown. The model of rate heterogeneity for each gene is shown, along with the number of generations required, and the number of markov chains (these values vary based on the size of the dataset).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S8.DOC]

Additional file 9: Additional Table 9 - Likelihood ratio tests (LRTs) performed using all evolutionary models used in selection analysis. Details on all likelihood ratio tests performed in the analysis. The models are denoted by their abbreviated names, Model A1 is denoted as Model A null throughout the manuscript. The number of degrees of freedom (df) are shown, this is relevant for the chi-squared test for significance, the critical values in each instance are given in the final column.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S9.DOC]

Abbreviations

A.A.: Amino Acid; Bck: Background lineage/s; BEB: Bayes Empirical Bayes; CDS: Coding DNA sequence; Dn: Non-synonymous substitution per non-synonymous site; Ds: Synonymous substitution per synonymous site; F: Frequency of amino acids; Fwd: Foreground lineage/s; G: gamma distributed sites rates across sites; GTR: General Time Reversible; I: invariable; JTT: Jones, Taylor and Thornton; LBA: Long Branch Attraction; LM: Likelihood mapping; LRT: Likelihood Ratio Test; ML: Maximum Likelihood; MSA: Multiple Sequence Alignment; N/A: data not available; NB: Non-binary tree; NEB: Naïve Empirical Bayes; NS: No statistical difference; OI: Osteogenesis imperfecta; OI-II/-III/-IV: Osteogenesis imperfecta type -2/-3/-4; P: probability; PP: Posterior Probability; RMSD: Root Mean Squared Deviation; SH: Shimodaira Hasegawa; SNP: Single nucleotide polymorphism.

Acknowledgements

We would like to thank the Irish Research Council for Science, Engineering and Technology (Embark Initiative Postgraduate Scholarship to NBL and CCM) for financial support and DCU School of Biotechnology scholarship (for TAW). We would like to thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for processor time and technical support for both phylogeny reconstruction and selection analysis. We would like to thank the SCI-SYM centre for processor time.

Authors' contributions

CCM carried out all data assembly, including searches of (i) literature, (ii) microarray studies, and (iii) protein interaction databases. CCM carried out all homolog identification and MSAs. NBL and CCM carried out all data quality and phylogeny analyses. TAW designed and performed randomization tests, designed bespoke software for the analyses and contributed to the preparation of the manuscript. CCM, NBL and MJO'C carried out all selective pressure analyses. NBL and CCM participated in drafting the manuscript. AJH analysed reproductive age data and genestational times for all mammals in the study, and helped to draft the manuscript. MJO'C conceived of the study, its design and coordination and drafted the manuscript. All authors read and approved the final draft.

Received: 20 July 2009
Accepted: 11 February 2010 Published: 11 February 2010

References

1. Aagaard JE, *et al*: **Rapidly evolving zona pellucida domain proteins are a major component of the vitelline envelope of abalone eggs.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103(46)**:1730-17307.
2. Wyckoff GJ, Wang W, Wu CI: **Rapid evolution of male reproductive genes in the descent of man.** *Nature* 2000, **403(6767)**:304-309.
3. McInerney JO: **The causes of protein evolutionary rate variation.** *Trends Ecol Evol* 2006, **21(5)**:230-2.
4. Zhou T, Drummond DA, Wilke CO: **Contact density affects protein evolutionary rate from bacteria to animals.** *J Mol Evol* 2008, **66(4)**:395-404.
5. Li WH, *et al*: **Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis.** *Mol Phylogenet Evol* 1996, **5(1)**:182-7.
6. Gaut BS, *et al*: **Relative rates of nucleotide substitution at the rbcL locus of monocotyledonous plants.** *J Mol Evol* 1992, **35(4)**:292-303.
7. Ohta T: **An examination of the generation-time effect on molecular evolution.** *Proc Natl Acad Sci USA* 1993, **90(22)**:10676-80.
8. Swanson WJ, Vacquier VD: **The rapid evolution of reproductive proteins.** *Nature reviews Genetics* 2002, **3(2)**:137-144.
9. Dorus S, *et al*: **Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity.** *Nature genetics* 2004, **36(12)**:1326-1329.
10. Anisimova M, Bielawski JP, Yang Z: **Accuracy and power of bayes prediction of amino acid sites under positive selection.** *Molecular biology and evolution* 2002, **19(6)**:950-958.
11. Shyamsundar R, *et al*: **A DNA microarray survey of gene expression in normal human tissues.** *Genome biology* 2005, **6(3)**:R22.
12. He Z, *et al*: **Expression of Col1a1, Col1a2 and procollagen I in germ cells of immature and adult mouse testis.** *Reproduction* 2005, **130(3)**:333-41.
13. Murphy WJ, *et al*: **Resolution of the early placental mammal radiation using Bayesian phylogenetics.** *Science* 2001, **294(5550)**:2348-51.
14. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17(12)**:1246-7.
15. Schmidt HA, *et al*: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics (Oxford, England)* 2002, **18(3)**:502-504.
16. Strimmer K, von Haeseler A: **Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment.** *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94(13)**:6815-6819.
17. Loughran NB, *et al*: **The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions.** *BMC evolutionary biology* 2008, **8**:101.
18. Puigbo P, Garcia-Vallve S, McInerney JO: **TOPD/FMTS: a new software to compare phylogenetic trees.** *Bioinformatics (Oxford, England)* 2007, **23(12)**:1556-1558.
19. Behera MA, *et al*: **Thrombospondin-1 and thrombospondin-2 mRNA and TSP-1 and TSP-2 protein expression in uterine fibroids and correlation to the genes COL1A1 and COL3A1 and to the collagen cross-link hydroxyproline.** *Reproductive sciences (Thousand Oaks, Calif)* 2007, **14(8 Suppl)**:63-76.
20. Marini JC, *et al*: **Consortium for osteogenesis imperfecta mutations in the helical domain of type I collagen: regions rich in lethal mutations align with collagen binding sites for integrins and proteoglycans.** *Human mutation* 2007, **28(3)**:209-221.
21. Oyen O, *et al*: **Human testis cDNA for the regulatory subunit RII alpha of cAMP-dependent protein kinase encodes an alternate amino-terminal region.** *FEBS letters* 1989, **246(1-2)**:57-64.
22. Leclerc P, de Lamirande E, Gagnon C: **Cyclic adenosine 3',5'monophosphate-dependent regulation of protein tyrosine phosphorylation in relation to human sperm capacitation and motility.** *Biology of reproduction* 1996, **55(3)**:684-692.
23. Hunnicutt GR, Primakoff P, Myles DG: **Sperm surface protein PH-20 is bifunctional: one activity is a hyaluronidase and a second, distinct activity is required in secondary sperm-zona binding.** *Biology of reproduction* 1996, **55(1)**:80-86.
24. Swanson WJ, Nielsen R, Yang Q: **Pervasive adaptive evolution in mammalian fertilization proteins.** *Molecular biology and evolution* 2003, **20(1)**:18-20.
25. Arming S, *et al*: **In vitro mutagenesis of PH-20 hyaluronidase from human sperm.** *European journal of biochemistry/FEBS* 1997, **247(3)**:810-814.
26. Phelps BM, Myles DG: **The guinea pig sperm plasma membrane protein, PH-20, reaches the surface via two transport pathways and becomes localized to a domain after an initial uniform distribution.** *Developmental biology* 1987, **123(1)**:63-72.
27. Bleil JD, Wassarman PM: **Identification of a ZP3-binding protein on acrosome-intact mouse sperm by photoaffinity crosslinking.** *Proceedings of the National Academy of Sciences of the United States of America* 1990, **87(14)**:5563-5567.

28. Kim KS, Cha MC, Gerton GL: **Mouse sperm protein sp56 is a component of the acrosomal matrix.** *Biology of reproduction* 2001, **64(1)**:36-43.
29. Sherry ST, Ward M, Sirotkin K: **dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation.** *Genome Res* 1999, **9(8)**:677-9.
30. Gupta SK, *et al*: **Structural and functional attributes of zona pellucida glycoproteins.** *Society of Reproduction and Fertility supplement* 2007, **63**:203-216.
31. Swanson WJ, *et al*: **Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98(5)**:2509-2514.
32. Chakravarty S, *et al*: **Relevance of glycosylation of human zona pellucida glycoproteins for their binding to capacitated human spermatozoa and subsequent induction of acrosomal exocytosis.** *Molecular reproduction and development* 2008, **75(1)**:75-88.
33. Shabanowitz RB, O'Rand MG: **Characterization of the human zona pellucida from fertilized and unfertilized eggs.** *Journal of reproduction and fertility* 1988, **82(1)**:151-161.
34. Wassarman PM: **Mammalian fertilization: molecular aspects of gamete adhesion, exocytosis, and fusion.** *Cell* 1999, **96(2)**:175-183.
35. Patrat C, *et al*: **Zona pellucida from fertilised human oocytes induces a voltage-dependent calcium influx and the acrosome reaction in spermatozoa, but cannot be penetrated by sperm.** *BMC developmental biology* 2006, **6**:59.
36. Primakoff P, Hyatt H, Tredick-Kline J: **Identification and purification of a sperm surface protein with a potential role in sperm-egg membrane fusion.** *The Journal of cell biology* 1987, **104(1)**:141-149.
37. Evans JP: **The molecular basis of sperm-oocyte membrane interactions during mammalian fertilization.** *Human reproduction update* 2002, **8(4)**:297-311.
38. Wong GE, *et al*: **Analysis of fertilin alpha (ADAM1)-mediated sperm-egg cell adhesion during fertilization and identification of an adhesion-mediating sequence in the disintegrin-like domain.** *The Journal of biological chemistry* 2001, **276(27)**:24937-24945.
39. Iba K, *et al*: **The cysteine-rich domain of human ADAM 12 supports cell adhesion through syndecans and triggers signaling events that lead to beta1 integrin-dependent cell spreading.** *The Journal of cell biology* 2000, **149(5)**:1143-1156.
40. Carlson AE, *et al*: **CatSper1 required for evoked Ca2+ entry and control of flagellar function in sperm.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(25)**:14864-14868.
41. Podlaha O, Zhang J: **Positive selection on protein-length in the evolution of a primate sperm ion channel.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(21)**:12241-12246.
42. Avenarius MR, *et al*: **Human male infertility caused by mutations in the CATSPER1 channel protein.** *American Journal of Human Genetics* 2009, **84(4)**:505-510.
43. Podlaha O, *et al*: **Positive selection for indel substitutions in the rodent sperm protein catsper1.** *Molecular biology and evolution* 2005, **22(9)**:1845-1852.
44. Peter A, *et al*: **Semenogelin I and semenogelin II, the major gel-forming proteins in human semen, are substrates for transglutaminase.** *European journal of biochemistry/FEBS* 1998, **252(2)**:216-221.
45. Hurle B, *et al*: **Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage.** *Genome research* 2007, **17(3)**:276-286.
46. Chamary JV, Hurst LD: **The price of silent mutations.** *Sci Am* 2009, **300(6)**:46-53.
47. Hurst LD, Pal C: **Evidence for purifying selection acting on silent sites in BRCA1.** *Trends Genet* 2001, **17(2)**:62-5.
48. iHOP: **The iHOP database.**.
49. Ensembl: **Ensembl.**http://www.ensembl.org, cited.
50. Chenna R, *et al*: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic acids research* 2003, **31(13)**:3497-3500.
51. Rambaut A: **Se-AL Sequence alignment editor.** *Oxford* 1996.
52. Lanave C, *et al*: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20(1)**:86-93.
53. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics (Oxford, England)* 2003, **19(12)**:1572-1574.
54. Keane TM, *et al*: **Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified.** *BMC evolutionary biology* 2006, **6**:29.
55. Brinkmann H, Philippe H: **Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies.** *Mol Biol Evol* 1999, **16(6)**:817-25.
56. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19(12)**:1572-4.
57. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Computer applications in the biosciences: CABIOS* 1997, **13(5)**:555-556.
58. Yang ZW, Wong S, Nielsen R: **Bayes empirical bayes inference of amino acid sites under positive selection.** *Molecular biology and evolution* 2005, **22(4)**:1107-1118.

doi:10.1186/1471-2148-10-39
Cite this article as: Morgan *et al.*: Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins. *BMC Evolutionary Biology* 2010 **10**:39.

BMC Evolutionary Biology

BioMed Central

Research article

Open Access

The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions

Noeleen B Loughran[1], Brendan O'Connor[2], Ciarán Ó'Fágáin[2] and Mary J O'Connell*[1,2]

Address: [1]Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland and [2]School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland

Email: Noeleen B Loughran - noeleen.loughran@gmail.com; Brendan O'Connor - brendan.oconnor@dcu.ie; Ciarán Ó'Fágáin - ciaran.fagan@dcu.ie; Mary J O'Connell* - mary.oconnell@dcu.ie

* Corresponding author

Published: 27 March 2008

BMC Evolutionary Biology 2008, **8**:101 doi:10.1186/1471-2148-8-101

Received: 29 August 2007
Accepted: 27 March 2008

This article is available from: http://www.biomedcentral.com/1471-2148/8/101

Abstract

Background: The mammalian heme peroxidases (MHPs) are a medically important group of enzymes. Included in this group are myeloperoxidase, eosinophil peroxidase, lactoperoxidase, and thyroid peroxidase. These enzymes are associated with such diverse diseases as asthma, Alzheimer's disease and inflammatory vascular disease. Despite much effort to elucidate a clearer understanding of the function of the 4 major groups of this multigene family, we still do not have a clear understanding of their relationships to each other.

Results: Sufficient signal exists for the resolution of the evolutionary relationships of this family of enzymes. We demonstrate, using a root mean squared deviation statistic, how the removal of the fastest evolving sites aids in the minimisation of the effect of long branch attraction and the generation of a highly supported phylogeny. Based on this phylogeny we have pinpointed the amino acid positions that have most likely contributed to the diverse functions of these enzymes. Many of these residues are in close proximity to sites implicated in protein misfolding, loss of function or disease.

Conclusion: Our analysis of all available genomic sequence data for the MHPs from all available completed mammalian genomes, involved sophisticated methods of phylogeny reconstruction and data treatment. Our study has (i) fully resolved the phylogeny of the MHPs and the subsequent pattern of gene duplication, and (ii), we have detected amino acids under positive selection that have most likely contributed to the observed functional shifts in each type of MHP.

Background

Heme peroxidases are readily abundant enzymes that can be classified into two major families, namely the animal and non-animal peroxidases, that have arisen from two independent evolutionary events [1]. The non-animal peroxidases include plant, bacterial, fungal and protist [1]. The classical peroxidase cycle involves the reaction sequence from native enzyme through compound I, then compound II and finally back to native enzyme [2]. An alternative and highly important pathway that mammalian heme peroxidases (MHPs) pass through, depending on substrate availability, is the halogenation cycle [3]. In

the presence of H_2O_2 and a halide (especially iodide), myeloperoxidase (MPO) can catalyse a halogenation reaction that plays an important role in the antibacterial activity of leukocytes [4]. Animal peroxidases are a medically important group of enzymes implicated in many different diseases including asthma [5], Alzheimer's disease (AD) [6] and inflammatory vascular disease [7]. From biochemical studies it is believed that the heme peroxidases for mammals arose following a number of gene duplication events [3,8,9].

Gene duplication provides the raw material for evolution of diversity and is believed to be the principal source of new genes [10]. The process of gene duplication has a number of alternative outcomes, and remains a controversial issue. Gene duplicates may become functionally redundant [11], or functionally divergent. There are a number of ways in which functional redundant duplicates can be preserved [12,13]. It has been proposed that the preservation of duplicates can be brought about by degenerative mutations in the regulatory elements of the duplicates, this is referred to as the Duplication-Degeneration-Complementation model (DDC) [13]. The DDC model does not allow a role for positive selection in the evolution of duplicates and is based solely on a neutral model with degenerate mutations and subsequent negative selection. Under this model duplicates are preserved as each accumulates degenerate mutations, resulting in specific subfunctions that *in toto* ensure optimal fitness [13].

An alternative mode of duplicate retention is positive selection. For example, in direct contrast to the predictions of the DDC model it has been shown for human and mouse that the number of retentions and losses of duplicates fits more consistently with a model incorporating positive selection [14]. Rapid divergence in gene expression profiles of duplicates following the duplication event results in expression profiles as diverse as those of singletons. An example of this is the functional redundancy of transcription factor inhibitors, Iκα and β, that have acquired different functions through divergence of gene expression rather than biochemical function [15]. Recent studies have indicated that for mammalian genomes neofunctionalisation, be it independent of -, or coupled with – subfunctionalisation, is the most common mode of evolution of gene duplicates [16]. These selective pressures following the process of gene duplication are key to the evolution of specificity of divergent multigene families, such as the MHPs [17].

In those cases where having all duplicates is deleterious, dosage requirements may cause the partitioning of subfunctions to be favored by positive selection resulting from selective pressure for the fixation of nonfunctional or subfunctional alleles. The divergence of function may occur through neofunctionalisation [18], or, subfunctionalisation where the ancestral function is partitioned between the duplicates [19] (for detail on current gene duplication models see [20]).

We hypothesise that the selective pressures on MHPs following gene duplication events will, (i) still be traceable in the extant sequences of these enzymes, and (ii), will have contributed to the functional diversity observed in these enzymes. A fully resolved phylogeny can provide a basis for such comparative genomic analysis of these heme peroxidases.

MHPs have been classified into four main families based on their function; myeloperoxidase (MPO), eosinophil peroxidase (EPO), lactoperoxidase (LPO) and thyroid peroxidase (TPO). MPO, EPO and LPO function in antimicrobial and innate immune responses [21-23], whereas, TPO plays a key role in thyroid hormone biosynthesis [24], see Table 1. A study of the structure-function relationships of human heme peroxidases suggest that the evolution of TPO succeeded that of MPO, EPO and LPO, but that these families shared a common ancestor [3,8,9]. MHPs are present in various tissues and as such their peroxidase function varies depending on tissue of expression. There are both structural and functional similarities among this multigene family of enzymes particularly with respect to their catalytic domains, this reflects their evolutionary relatedness. It has been shown that active site residues are conserved in all heme peroxidases [3,25].

To infer the phylogeny of the MHPs from sequence data, it is fundamental to consider the challenges associated with resolving mammalian gene phylogenies. The main pitfalls include poor phylogenetic signal resulting from mutationally saturated positions, inadequate modelling of the evolutionary process and systematic bias due to var-

Table 1: Mammalian heme peroxidase features and functions (adapted from Clark 2000 and O'Brien 2000).

Superfamily (EC no.)	Chromosomal Location (Human)	Tissue Expression	Biological Function
MPO (1.11.1.7)	17	Neutrophils, mono-nuclear phagocytes	Microbicidal activity
EPO (1.11.1.7)	17	Eosinophils	Microbicidal activity
LPO (1.11.1.7)	17	Milk, saliva, tears and other secretions	Bacteriostatic and bactericidal activity
TPO (1.11.1.8)	2	Thyroid cell surface and cytoplasm	Thyroid hormone biosynthesis

MPO = Myeloperoxidase; EPO = Eosinophil peroxidase; LPO = Lactoperoxidase; TPO = Thyroid peroxidase.

iable rates of evolution among species or within sequences [26].

A systematic bias or systematic error is one that results in greater support for an incorrect conclusion with the accumulation of more data. Long branch attraction (LBA) is one of the most commonly occurring systematic biases and is a consequence of unequal evolutionary rates across lineages. This can occur due to the number of cell divisions per unit time being different in different species or due to rapid fixation of mutations due to reduced population size, e.g., a bottleneck. Rodent species accumulate many more mutations within a defined time frame than larger mammals [27,28]. Therefore, rodentia are often placed close to the outgroup species on a phylogeny due to their increased number of mutations. There are a number of ways in which the noise (LBA) can be minimised. Firstly, the addition of more taxa to the dataset: denser sampling of species of intermediate generation time can reduce the effect of LBA by reducing the overall distances between taxa. Secondly, the use of improved models of sequence evolution, i.e., models sensitive to multiple substitutions at the same site and rate heterogeneity across the phylogeny. And finally, stripping the alignment of its most rapidly evolving sites and using only the remaining more slowly evolving sites to reconstruct phylogenies reduces the amount of LBA noise in the dataset [29]. These approaches can be used in combination. While databases such as Peroxibase [30] house all the up-to-date peroxidase sequences [31], we have included only those MHPs from completed mammalian genomes (allows us identify species-specific gene birth and death). We have used Maximum Likelihood (ML) and Bayesian methods of phylogeny reconstruction together with the stripping of the most rapidly evolving sites in the dataset.

The major questions addressed in this study pertain firstly to the resolution of the evolutionary relationships of these MHPs using molecular sequence data, and secondly, to the analysis of functional diversities among these superfamilies using the resolved phylogeny and ML methods for testing selective pressures.

Selection can be classified as being neutral, purifying or positive. Positive selection/Adaptive evolution is strongly indicative of functional shifts within proteins [32]. To determine what selective pressures may have influenced the functional diversification of the MHP families, we tested the data using a variety of ML models of evolution with different properties. These included models that allow for only purifying selection and/or neutral evolution, and those that allow for positive selection. Likelihood scores for all alternative models and their null hypotheses are calculated. The likelihood scores for the null hypothesis versus the alternative hypothesis for those models that are extensions of each other were then compared using a likelihood ratio test (LRT) for goodness-of-fit. For those models that allow for the estimation of site-specific evolution, we can identify those amino acids that have undergone positive selection. The location of these amino acid positions were estimated using Bayesian statistics and their location and possible functional significance were determined. In our analysis we have shown that positive selection has contributed to the evolution of these enzymes following gene duplication events.

Results

Phylogeny Reconstruction

The MHP dataset for this study consisted of 31 single gene orthologues from MPO, EPO, LPO, and TPO classes, totaling 1,017 aligned positions. The species phylogeny for the mammals has previously been fully resolved [33]. In brief, the mammalian species phylogeny describes Marsupialia (i.e. Opossum in our dataset) as outgroup to all other mammals, followed by the divergence of the Carnivora (i.e. Dog in our dataset) and the Cetacea (i.e. Cows in our dataset), and finally the emergence of the Euarchontoglires clade (i.e. primates and rodents) [33], see Figure 1a. The ML phylogenetic tree was estimated using MultiPhyl [34] and MrBayes 3.1.2 [35], the results were congruent, see Figure 2a. Each of the four superfamilies branched into their respective functional groups, with the members of the TPO superfamily taking the position of outgroup with high support values. The topology shows MPO, EPO and LPO shared a most recent common ancestor (MRCA) with a gene duplicate of TPO. The MPO and EPO groups themselves shared a MRCA and functionally diverged following a further gene duplication event. Therefore these two peroxidases (MPO & EPO) are the most closely related of all the MHPs in this study.

Despite the 4 major clades in the phylogeny corresponding to the 4 major groups of MHPs, the relationships of the species within these clades conflicts with the previously published mammalian species phylogeny [33]. The rat and mouse are members of the glires group, and as such are a sister group to the primates, which together form the Euarchontoglires mammalian superorder. The topology seen here for the LPOs (see Figure 2a) suggests that dog and cow are the outgroup to the primate clade. This is a common error in mammalian phylogeny reconstruction, and has been proven to be an effect of LBA [36]. Also, for the TPO group opossum is placed next to rat and mouse and not as the outgroup as expected, suggesting that the opossum and the rodents have similar rapid rates of evolution, see Figure 2a.

We adapted the site stripping method using the slow-evolving positions for each species in the MSA to reconstruct the phylogeny, while still retaining adequate

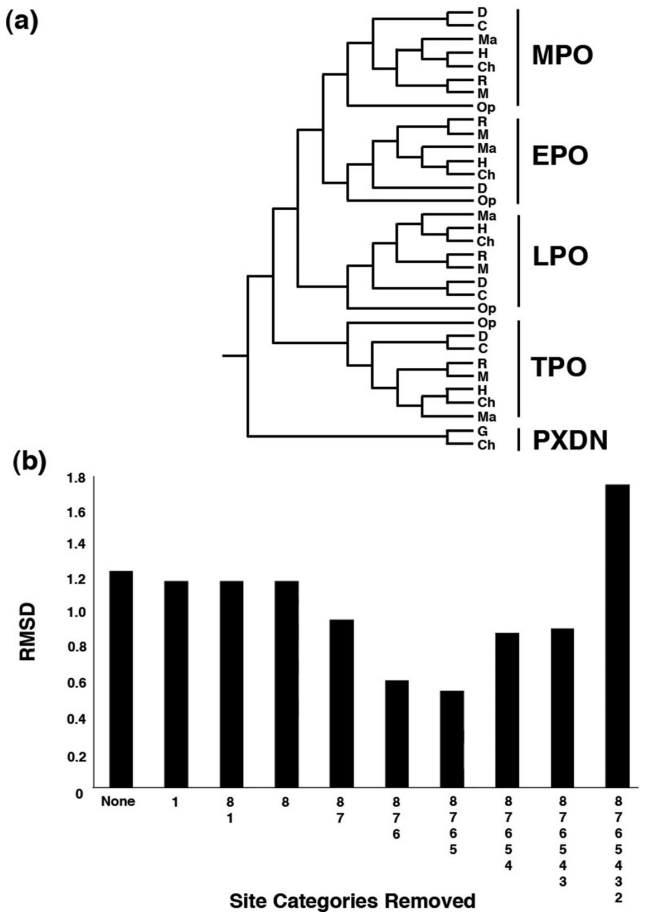


Figure 1
The distance between each of the site stripped phylogenies and the ideal mammalian peroxidase phylogeny. **(a)** The ideal phylogeny pruned from the mammalian phylogeny by Murphy *et al.* (2001), the peroxidasin sequences are outgroups to the MHP clade. The following are the species abbreviations used: Dog (D); Cow (C); Macaque (Ma); Human (H); Chimp (Ch); Rat (R); Mouse (M), Chicken (G), and Opossum (Op). This phylogeny was compared to each of the resultant site stripped phylogenies. **(b)** Graph showing the RMSD nodal distance (*y-axis*) between each site-stripped phylogeny (*x-axis*) and the ideal phylogeny. *On the X axis*: All: refers to the complete MSA; 8: site category 8 removed from the MSA; 8, 7: categories 8 and 7 removed from the MSA and so on up to the final column that contains only the most slowly evolving category of site. Values close to/zero correspond to complete agreement between the ideal and site stripped phylogeny.

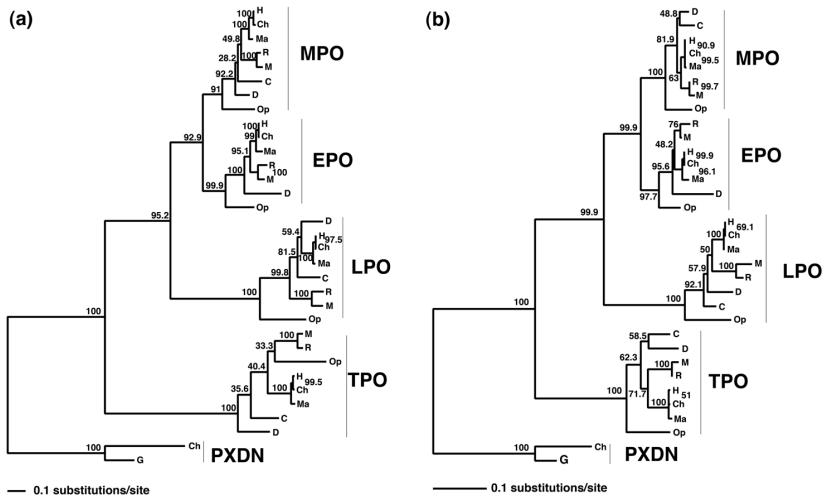


Figure 2
Phylogeny of the mammalian heme peroxidases before treatment for long branch attraction and after treatment. (a) Initial unresolved ML tree for mammalian heme peroxidases and peroxidasin from *Pan troglodytes* and *Gallus gallus* from the entire dataset. The bootstrap support values from 1000 replicates are shown on all nodes. (b) Resolved phylogeny following site stripping, the cow sequence for LPO can be seen to take an unusual place on the phylogeny.

amounts of signal [29]. This approach is similar to the '*Slow-Fast Method*' [37] and is therefore an approximate method that removes noise from the data by removing those sites that are most likely to contain homoplasy and focusing on the more evolutionary informative positions for phylogeny reconstruction. Each site within the MSA was classified according to rates of evolution (estimated using ML based on a fixed phylogenetic tree). To determine what number of categories to remove, we progressively stripped each category from the most rapidly evolving sites to the most slowly across the entire MSA. We also combined removal of the fastest and slowest sites from the dataset in our analysis, this was initially performed with the PXDN data included, see Figure 1b. Each time a category was removed the phylogenetic tree was estimated from the remaining MSA using ML. The ideal tree was created by pruning the mammalian supertree as published by Murphy *et al.* [33] (with the inclusion of chicken) and is depicted in Figure 1a. The difference between each site-stripped phylogeny and the ideal phylogeny was calculated using a nodal distance calculation RMSD [38], see Figure 1b. From Figure 1b, it is seen that the removal of rapidly evolving sites gradually removes the noise from the data and the remaining signal moves towards the canonical species phylogeny [33]. For the dataset consisting of MHPs and PXDN sequences, the RMSD value reaches a minimum at the removal of 4 site categories (8, 7, 6 and 5) leaving a MSA of length 850 sites (including gaps/missing data), see Figure 2b for resultant topology, after this point the RMSD values rise, see Figure 1b. It is important to note that the slowest evolving positions can be misleading particularly with excessive removal of sites, as the number of characters for reconstruction will decrease with every cycle, therefore caution must be taken in applying this method. This analysis was also performed on the dataset containing only MHP sequences, and the RMSD value reaches a minimum at the removal of 3 site categories (8,7, and 6) leaving a MSA of length 613 sites (including gaps/missing data), see Figure 3a for resultant topology. The reduced MSA for MHP data is given in Additional file 1 and the corresponding TOPD results are given in Additional file 2. The nodal distance (RMSD) calculation is based entirely on the branching pattern and hence does not account for evolutionary rate variation across the phylogeny. Using this site-stripped MSA the phylogeny was estimated using both MrBayes

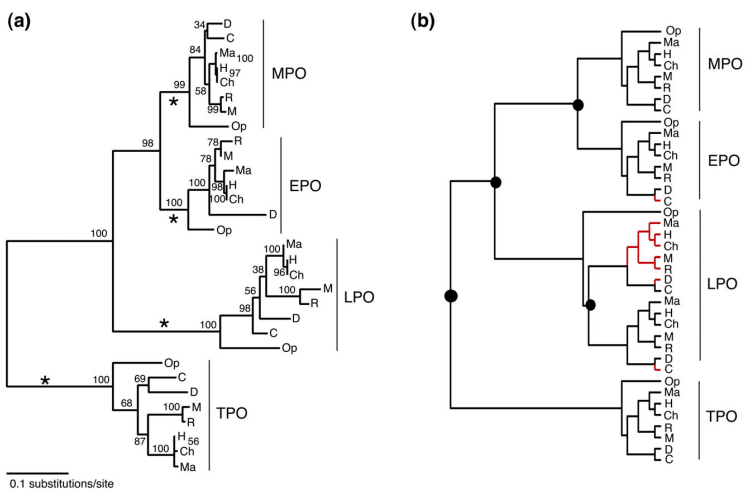


Figure 3
Fully resolved mammalian heme peroxidase phylogeny with duplication and loss events depicted. **(a)** Resolved ML tree for mammalian heme peroxidases. The bootstrap support values from 1000 replicates are shown on all nodes. The TPO primate clade appears here as a polytomy as the branch lengths are extremely short, however, this is in fact resolved with a low Bootstrap of 56%. The star symbol denotes those branches that were treated as foreground in the selection analysis. **(b)** The analysis of the resolved phylogeny using gene tree species tree reconciliation method implemented in GeneTree. The large filled circles represent gene duplication events, and the red branches indicate gene losses.

and MultiPhyl methods, both of which produced identical phylogenies*. (*We note here that the one exception, using the Bayesian reconstruction method, was the TPO primate monophyly was not fully resolved in the TPO clade but instead supported a human-chimp-macaque polytomy.)

All gene duplication events were verified using gene tree – species tree reconciliation. We analysed the resolved MHP phylogeny (Figure 3a), and identified in total 4 duplication events and 4 losses. This method over prescribes gene losses as in the case of EPO, where the sequence data was not available and therefore is assumed to be a loss. There is an LPO specific duplication event predicted, see Figure 3b. Our results show differential retention and loss in the LPO lineage following this gene duplication event resulting in the cow species retaining an alternative duplicate copy to the other mammals in the dataset, as shown in Figure 3b. This method must be used with caution as it does not take into account rate heterogeneity amongst species or sites in the data, and relies solely on the topology. However, reciprocal BLAST analysis of the cow sequence against the other mammal genomes identifies this sequence as an ortholog.

Functional Diversity and Evolution of Specificity

We wished to test the hypothesis that following the gene duplication events in the MHPs (as resolved in this study), selective forces – specifically positive selection – have contributed to the observed changes in function in each of the 4 major groups of MHPs. Tests for heterogeneous selective pressures were carried out on the resolved phylogeny using the evolutionary models implemented in PAML 3.15 [39] and the complete MSA. The Dn/Ds ratios were estimated in a likelihood framework at both site-specific and lineage-specific levels. A total of seven tests of significance were carried out using χ^2 tests of significance, five site-specific comparisons and two branch-site comparisons were performed.

No positively selected sites were estimated for the one ratio model (see Additional file 3). Strong purifying selec-

tion across sites was indicated with an ω of 0.1516. However, this model is a poor fit for the data (ln*L* = -34417.1085). Positive selection was tested in a site-specific manner across the dataset using the site models; M1 (neutral), M2 (selection), M3 discrete (k = 2), M3 discrete (k = 3), M7 (beta), M8 (beta & omega > 1) and M8a (beta & omega = 1). The results of the site-specific analysis are shown in Additional file 3.

Poor likelihood values were achieved using the site-specific models of evolution, however, the most complex site-specific model used, M8 yielded significant results when it was tested with its null model M8a. A small proportion of sites are under relaxed positive selection (Additional file 3). Through the use of Bayesian estimations, four positively selected sites have been identified across the alignment, with posterior probability (PP) > 0.50.

Results of the branch-site model B with each of the families individually labeled as foreground are shown here in Table 2; see Figure 3a for corresponding foreground branches. (Results for model A are given in Additional files 4 and 5). To determine whether there is rate heterogeneity along different branches in the phylogeny, we compared models allowing for only site-specific evolution with those allowing for branch-site specific evolution (i.e. M3 K = 2 with Model B and M1 with Model A). Following LRT analysis it was found that both models A and B were significant following χ^2 test with two degrees of freedom. The likelihood score from model B for each family had improved significantly from those obtained using model A, as a result, model B was determined as the best fit model in each case tested and these results are summarized in Table 2. Positively selected sites identified with model B were estimated using the Naïve Empirical Bayes (NEB) method [40]. The results of which are discussed now in detail.

Our results show that following gene duplication, each individual type of MHP has undergone positive selection in amino acid residues that are unique to that type of MHP, see Table 2. As positive selection is closely associated with functional shift, we postulate that these positively selected sites have significantly contributed to the evolution of the functional diversity of these MHPs.

For the MPO superfamily, a total of 19 positively selected sites were identified (PP > 0.50). We have found functional information from the literature on 11 of these sites, these are now discussed: Position 80 (Arg) is located within the propeptide sequence and is under positive selection. Previous studies indicate that propeptide in MPO plays a key role in the processing and sorting of human MPO [41]. Position 568 is under positive selection and is next to the polymorphic site R569W, mutations in position 569 have been shown to suppress posttranslational processing in MPO [42]. The 2 positions with strongest support, PP > 0.95, are separated by 8 amino acid residues on the MPO heavy chain, they are Asn496 and Leu504. These 2 positions along with Tyr500

Table 2: Parameter estimates and likelihood scores for branch-site model, model B.

Model	P	L	Estimates of parameters	Positively selected sites
MPO Model B	5	-33655.0405	p_0 = 0.4975, p_1 = 0.4553, (p_2 = 0.0246, p_3 = 0.0225) *Background*: ω_0 = 0.0458, ω_1 = 0.3307, ω_2 = 0.0458, ω_3 = 0.3307 *Foreground*: ω_0 = 0.0458, ω_1 = 0.3307, ω_2 = 251.6783, ω_3 = 251.6783	**Foreground**: NEB 19 > 0.50 2 > 0.95 1 > 0.99
EPO Model B	5	-33647.5634	p_0 = 0.4967 p_1 = 0.4469, (p_2 = 0.0297, p_3 = 0.0267) *Background*: ω_0 = 0.0464, ω_1 = 0.3322, ω_2 = 0.0464, ω_3 = 0.3322 *Foreground*: ω_0 = 0.0464, ω_1 = 0.3322, ω_2 = 774.6323, ω_3 = 774.6323	**Foreground**: NEB 28 > 0.50 6 > 0.95 4 > 0.99
LPO Model B	5	-33627.3508	p_0 = 0.4431, p_1 = 0.3884, (p_2 = 0.0898, p_3 = 0.0787) *Background*: ω_0 = 0.0470, ω_1 = 0.3414, ω_2 = 0.0470, ω_3 = 0.3414 *Foreground*: ω_0 = 0.0470, ω_1 = 0.3414, ω_2 = 82.8559, ω_3 = 82.8559	**Foreground**: NEB 96 > 0.50 18 > 0.95 11 > 0.99
TPO Model B	5	-33639.5793	p_0 = 0.4358, p_1 = 0.3690, (p_2 = 0.1057, p_3 = 0.0895) *Background*: ω_0 = 0.0479, ω_1 = 0.3468, ω_2 = 0.0479, ω_3 = 0.3468 *Foreground*: ω_0 = 0.0479, ω_1 = 0.3468, ω_2 = 999.0000, ω_3 = 999.0000	**Foreground**: NEB 82 > 0.50 8 > 0.95

Model B allows each foreground lineage to be tested independently of all other lineages, hence the four clusters (MPO, EPO, LPO, TPO – each in turn treated as foreground), and estimates 5 parameters (P) in total. p_0, p_1, p_2 and p_3 are proportions of sites in the dataset with the corresponding ω value, i.e, ω_0, ω_1, ω_2 and ω_3 for the foreground and the background lineages independently. The final column gives the estimated number of sites with posterior probabilities of greater than 0.50 of belonging to the positively selected category. Note: NEB: Naïve Empirical Bayes analysis.

are in close proximity to the proximal heme ligand in MPO, His502 [3]. Position 259 (Leu) is located between two important distal residues, Gln257 and His261, involved in the formation of hydrogen bonds [3]. His261 has an important role in the formation of compound I, a redox intermediate of the peroxidase cycle [2]. A further four sites (Leu630, Gln633, Glu652; (primates Lys652) and Asn654 (primates Lys654) were identified as positively selected, PP > 0.70, these are located within a disulfide bond linking helices 19 and 22 on the MPO heavy chain. Disulfide bonds are associated with the folding and stability of proteins and as such are significant to the overall function of that protein [43].

For the EPO clade, 28 sites are positively selected, PP > 0.50. We have found functional information for 15 of these sites. One of these, Asp71, is located in the EPO propeptide. The inferred phylogeny, shown in Figure 3a, suggests that MPO and EPO are closely related enzymes, therefore it may be possible that the EPO propeptide may also be crucial for the function of EPO. The region separating the catalytic residues Arg377 and His474 [3], contains 8 positively selected sites (PP > 0.50). Arg377 is the conserved prominent distal amino acid associated with hydrogen bond formation. The proximal heme ligands His474 (EPO), His502 (MPO) and His468 (LPO), are conserved in all the MHPs [3,25]. Six of the 28 positively selected sites, Arg584, Gln588, Arg591, Ala618, Gly626 and Ala627, are located on the EPO heavy chain within a single disulfide bond region, this would suggest that they are structurally and functionally important to EPO. Position 441 has been identified as under positive selection, this residue has also been noted as being polymorphic in the human population (Lys/Thr).

There are 18 positively selected sites for the LPO group (PP > 0.95). We have found functional information on 13 of these sites. Residues Glu72, Asn87 and Trp91 are found in the LPO propeptide sequence and have a probability of greater than 0.95 of being positively selected. Residues Asn255, Phe282, Ser312, Ser352 and Glu355 are all located in the disulfide bond region (PP > 0.95). From biochemical analysis both Arg372 (Arg377 in EPO) and His468 are believed to have catalytic properties, and are conserved in the MHPs [3,25]. We find positive selection in His376 (PP > 0.99) just four amino acids downstream of the first of these catalytic residues (Arg372), interestingly this site is specific to the primate lineage. Also we have detected positive selection in Glu470 (PP > 0.98) adjacent to the second catalytic site (His468). We have also detected positive selection in Asp700 which is a known genetic variant and Glu240 and Gln245 that are located to the right and left of a known human polymorphism A244T.

With the TPO clade treated as foreground, 8 sites are positively selected, PP > 0.95. Of these 8 sites, 6 are missing in the alternatively spliced TPO isoform 5, which exhibits incorrect protein folding [44]. Asp228 (PP > 0.95), Ala232 and Ala242 (both PP > 0.50) are in the region of the TPO active site His239. Glu378 has also been identified as a novel mutational site (E378K) associated with the common inherited deficiency total iodide organification defect (TIOD) and is under positive selection in our analysis [45].

Independent analysis for positive selection using DIVERGE [46] software further supports our findings, see Table 3 for summary of results. Values greater than zero for the coefficient of functional divergence, θ, indicate a functional shift between clusters. Rate heterogeneity among sites varies with respect to the gamma distribution (α). We estimated θ for each of the four MHP clusters. This analysis shows significant functional constraints among the four MHP clades, with the null hypothesis θ = 0 being rejected for all clusters analysed. The analysis of closely related MPO and EPO clusters result in the lowest θ value (0.2833 +/- 0.0837), and both have microbicidal activity (Table 1). θ increases at least 1.5 fold for the more distantly related/functionally divergent clusters. These results provide statistical evidence of the diverse functions of these MHP enzymes.

We further test the relationship between positive/directional selection and functional shift by analyzing the effect of these substitutions on the MPO 3D structure, see Figure 4. Modeling the MPO human sequence using SwissModel and using the mutate tool in DeepView v3.7, we have performed *in silico* site directed mutagenesis on those sites identified in our study as being positively

Table 3: Summary of results of analysis using DIVERGE software.

	MPO/EPO	MPO/LPO	MPO/TPO	EPO/LPO	EPO/TPO	LPO/TPO
θ **ML**	0.2832	0.4504	0.4984	0.4552	0.4304	0.4280
SE θ	0.0837	0.0744	0.0783	0.1021	0.0950	0.0756
LRT θ	11.4512	36.6860	40.4815	19.8713	20.5223	32.0448
α **ML**	0.3034	0.4221	0.4172	0.4863	0.4654	0.5413

Each cluster analysed is shown in the columns of the table. θ **ML**: Coefficient of functional divergence. **SE** θ: Standard error of the estimate Theta. **LRT** θ: 2 log-likelihood-ratio against the null hypothesis of θ = 0. α **ML**: Gamma shape parameter for rate variation among sites.

selected [47,48]. The structure with positively selected sites and the heme binding site is shown in Figure 4a. We find that mutating these positions from their positively selected state to the ancestral state causes a variety of effects on the hydrogen bond formation within the 3D structure, see Table 4 for a summary of the effects on hydrogen bonds. Hydrogen bonds play an important role in maintaining the structural integrity of a protein, any disruption of such forces is likely to upset the balance between the structural and functional dynamics [49]. On mutating each of these 19 positively selected amino acids we find that 4 bonds are lost and 4 are independently gained in the protein, for summary see Table 4. For the mutations: N496F, Y500F, and L504T, the positions of the losses and gains of hydrogen bonds are significant as these amino acid are in close proximity to the proximal heme ligand His502, shown in Figure 4a. The mutation from leucine to threonine at position 504 results in the formation of an additional hydrogen bond between Gly501 and Leu504. Gly501 is directly bound to the proximal heme ligand. In addition, the N496F mutation illustrated in Figure 4b, results in the loss of the hydrogen bond with Asn587. The Asn587 and His502 are connected by a hydrogen bond [3]. The loss of the hydrogen bond, as a result of the mutation at position 496, is likely to affect the structural integrity of the link between Asn587 and His502. Disruption to the hydrogen bonds in this catalytically important region may have direct implications for functional divergence of the MPO enzyme. The A471R mutation results in an increase in the number of hydrogen bonds associated with this position. This position is upstream from Asn483 which is thought to be responsible for MPOs dimer interaction [3]. The mutation from cysteine to serine at position 316 results in the formation of a hydrogen bond with Gln329 and the loss of one of the bonds to Asp593, see Table 4. Cys316 is next to the single disulphide bridge (Cys319) that connects MPOs symmetry-related halves [3]. The C316S mutation may potentially disrupt this disulphide bridge.

Table 4: Summary of results from SwissModel analysis of positively selected sites.

Mutation	Posterior Probability	Affect on Hydrogen Bond
C316S	0.815	-/+
S414A	0.600	-
A471R	0.738	+
P477G	0.948	=
N496F	0.999	-
Y500F	0.731	-
L504T	0.970	+
R529E	0.657	+
I568L	0.686	=
P584A	0.949	=
L630F	0.767	=
Q633L	0.737	=
L652V	0.840	=
L654G	0.921	=
S687T	0.648	+

Mutation from positively selected site in MPO (using human model) to the amino acid present in EPO, LPO and TPO at that position (in cases where there was conflict the majority rule consensus at that position was taken). Posterior Probability values extracted using NEB analysis in model B Codeml. Effect on H-Bonds is classified as "+" if an increase in the number of bonds with positively selected amino acid, "-" if a hydrogen bond or a number of hydrogen bonds were lost with the positively selected site, and "=" refers to no affect on the hydrogen bond with the positively selected site.

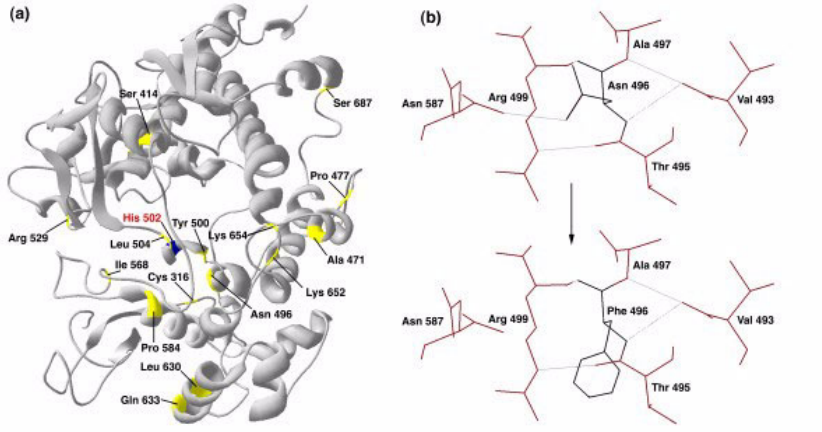


Figure 4
Location of positively selected sites in the MPO structure and their effect on bonding within the structure. (a) 3-D structure of the human MPO sequence, highlighted in gold are those sites that are positively selected in MPO, in blue is the heme binding site. **(b)** Example of the affect on hydrogen bonding of one such mutation at positively selected position 496 in human MPO from Asparagine to Phenylalanine.

Discussion and Conclusion

The MHPs are a functionally diverse family of enzymes which are implicated in a variety of inflammatory and neurodegenerative diseases such as asthma and AD respectively. In this study the evolutionary history of the four major groups of MHPs; MPO, EPO, LPO and TPO, was investigated allowing for the analysis of their functional diversity.

Initial ML and Bayesian phylogenies estimated here for the MHPs support previous biochemical studies [3,8,9]. From Figure 3 the order of gene duplication events can be traced, with an MPO-EPO-LPO MRCA arising from a gene duplication with extant TPO; then a further duplication event that gave rise to, (i) the MPO-EPO MRCA, and (ii), the lineage leading to extant LPO; and the final and most recent duplication of the MPO-EPO MRCA into extant MPO and EPO clades. PXDN is the outgroup to the MHP sequences and was included in the analysis to illustrate that TPO is the most ancestral MHP (Figure 2a). However, the species relationships estimated within these clearly defined clades were in disagreement with the previously resolved mammalian phylogeny [33].

Including all sites of the alignment in the analysis, we have shown that the major types of MHP form monophyletic clades and are therefore the result of gene duplication events prior to speciation of modern day mammals, see Figure 2(a). However, also evident from Figure 2(a), species with more similar generation times are clustered together, with species of shorter generation times and therefore more rapid rates of mutation assuming a basal position in the phylogeny. This observed branching pattern could be a result of LBA, incorrect ortholog prediction or hidden paralogy.

If a phylogeny is seen to approach the ideal by removing the most rapidly evolving sites, then we propose that LBA is most likely to have contributed to the misleading phylogeny. To test for the presence of LBA we calculated 8 categories of rates of evolution for all sites, from the most rapidly evolving to the most slowly evolving. We observed that the sequential removal of rapidly evolving categories of sites from the alignment decreased the difference, in terms of nodal distance RMSD, between the phylogeny produced and the ideal phylogeny. This occurred only for removal of the 4 fastest evolving categories of site from the alignment. Further removal after this point resulted in increased RMSD values between the phylogeny produced and the ideal. The MHP phylogeny shown in Figure 3(a), with maximum number of sites and minimum amount of noise. We propose that a possible reason for the presence of LBA in this dataset is the presence of taxa with vastly different generation times. The rodentia have previously been shown as "fast evolving" due to their short germ-line generation time, whereas species such as dogs and humans have longer germ-line generation times [27,28,50]. In any given dataset there are sites that are variable and sites that are invariable, this pattern is conserved across homologous sequences. In a dataset with a mixture of germ line generation times, the mutation rate in the species with shorter germ line generation times will be higher, because the number of cell divisions per unit time is greater. Therefore the number of mutations in the variable regions will increase for these species. The result is an LBA effect derived from having a mixture of long and short germ line generation times in the dataset, where the species with a short germ line generation time assumes a basal position in the phylogeny [26-28]. A number of approaches have been explored to systematically deal with fast evolving taxa the most popular include, (1) reconstructing the phylogeny based on slow evolving sites (applied here), (2) increasing the sample size, this is based on the assumption that increasing the sample size actually increases the number of slowly evolving positions, (3) decreasing the distance to the outgroup, and (4) using more accurate models of sequence change such as covarion derivatives.

Our gene tree – species tree reconciliation analysis has verified the duplication pattern amongst the MHPs. However, we believe that current methods of reconciliation such as the one used here may be biased towards inferring excess gene duplication and differential loss events, as is the case here. The method only considers the topology and not the corresponding alignment or any rate heterogeneity that may exist [51]. We would also like to highlight that the variation of the "*Slow-Fast*" method employed here is an approximate method for a complex evolutionary dynamic and is not without its limitations.

Using this fully resolved phylogeny, positively selected sites have been identified, through the use of Bayesian estimation, unique to all four MHPs; MPO, EPO, LPO and TPO. The majority of these sites are in close proximity to catalytically important residues, suggesting that they may

potentially be linked to functional shifts across the MHPs. The conserved proximal histidines in close proximity to sites under positive selection in MPO, EPO and LPO are crucial in preserving the redox properties of the heme iron for catalysis [3]. The conserved distal histidines, also shown here to be in the vicinity of positively selected sites, act as both proton acceptors and donor to oxygen during the formation of Compound 1, which is an integral step in the peroxidase pathway [3]. A number sites identified under positive selection are located in disulphide bond regions, which are believed to be crucial to the structure and function of a protein. Disruption of such regions can be detrimental to the enzymatic stability and activity [43,52]. In particular, six sites pertaining to the LPO family are linked to the same disulphide bond. This strongly suggests that these sites are associated with the unique function of LPO as they are not present in the two closely related families MPO and EPO. In the TPO analysis the majority of the sites with highest probability of being positively selected are located in exon 8 of the protein. Deletion of exon 8 results in misfolding of the TPO protein [44]. Exon 8 is also believed to be part of TPOs catalytic centre (exons 8, 9 and 10) [53]. TPO functional defects are strongly associated with TIOD and several deleterious mutations within this catalytic region have been reported [44,53-55]. We also find that one of our positively selected sites in TPO is associated directly with an inherited deficiency disorder [55].

Our detailed *in silico* site directed mutagenesis of the positively selected sites in MPO has shown that mutating these positions from their positively selected amino acid state to an alternative ancestral state results in loss/gain of hydrogen bonds between alternative amino acid positions for other sites in particular in the heme binding region of the MPO structure. The sites we have identified as positively selected in the MHPs have played a major role in the functioning of these enzymes as evidenced by mutational studies, proximity to active sites and catalytic residues, and inherited disorders.

The results of this study show for the first time from molecular sequence data (i) how this medically important group of enzymes are related to each other, and (ii) suggest that following gene duplication, positive selection has led to the functional diversity observed for the MHPs.

Methods

Sequence Data

Protein coding sequences for MHPs were retrieved from the Ensembl database for all available completed mammalian genomes using the pre-defined orthologues identified in Ensembl [56]. The mammalian genomes and the corresponding genome versions used for each of the major families in our dataset were as follows: *Homo sapiens* v42.36d; *Pan troglodytes* v42.21a; *Macaca mulatta* v42.10b; *Mus musculus* v42.36c; *Rattus norvegicus* v42.34l; *Canis familiaris* v42.2; *Bos taurus* v42.2e (no EPO sequence available), and, *Monodelphis domestica* v42.36c. Ensembl identifies orthologues by performing a genome-wide reciprocal WUBlastp+SmithWaterman search of each gene across all completed genomes. Multiple sequence alignment (MSA) is then performed using the MUSCLE software [57] and the best reciprocal hits following the sequence similarity search. The longest alternative transcript in each case was used. These sequences were combined into a single MHP dataset of 31 sequences. Two amino acid sequences representing the peroxidasin (PXDN) family, from the *Pan troglodytes* and the *Gallus gallus* genomes, were retrieved from the PeroxiBase database [31]. The sequence data are given in Table 5.

Multiple Sequence Alignment

Each protein coding sequence in the MHP dataset was translated to amino acid using in-house translation software. This protein sequence dataset and the two PXDN sequences were combined to give a dataset of 33 sequences (complete dataset). Both MHP and "complete" datasets were aligned in ClustalW 1.8 [58] independently using default parameter settings. The corresponding nucleotide sequences for the MHP dataset were aligned with respect to the amino acid MSA with the use of in-house software to insert gaps in the protein coding sequence according to their positions in the amino acid alignment. The nucleotide and subsequent protein MSAs were manually edited by removing ambiguous regions from the alignment using the sequence alignment editor, Se-Al 2.0a11 [59]. The PXDN sequences served as an outgroup for the MHPs and therefore aided in determining the earliest diverging MHP.

Site Stripping and Phylogeny Reconstruction

The phylogenetic tree for the dataset was estimated using Bayesian statistics implemented in MrBayes 3.1.2 [35]. The model of amino acid substitution used was JTT [60] because following model testing using MultiPhyl [34] this was the model that was best-fit to the data. Using 4 Markov chains for 400,000 generations, trees were sampled every 10 generations with the first 20,000 sampled trees discarded as burnin. The remaining trees samples were summarized on a majority rule consensus tree with clade supports given as Posterior Probabilities (PPs). ML trees were also inferred using the high-throughput phylogenomics webserver, MultiPhyl [34]. The ML tree was generated using the nearest neighbour interchange (NNI) tree search algorithm and 100 bootstrap replicates implemented in MultiPhyl [34] under the Akaike Information Criterion (AIC) statistic, the selected substitution model was JTT with invariable sites and a discrete gamma model of rate heterogeneity. This was repeated a total of 10 times

Table 5: Representative mammalian heme peroxidase sequences used in this study.

Superfamily	Species	Entry ID (Name)*/Gene ID	Length (aa)
MPO	*Homo sapiens*	ENSG00000005381	778
	Pan troglodytes	ENSPTRG00000009449	778
	Macaca mulatta	ENSMMUG00000002266	777
	Mus musculus	ENSMUSG00000009350	719
	Rattus norvegicus	ENSRNOG00000008310	719
	Canis familiaris	ENSCAFG00000017474	743
	Bos taurus	ENSBTAG00000012783	596
	Monodelphis domestica	ENSMODG00000014737	403
EPO	*Homo sapiens*	ENSG00000121053	716
	Pan troglodytes	ENSPTRG00000009446	716
	Macaca mulatta	ENSMMUG00000011973	717
	Mus musculus	ENSMUSG00000052234	717
	Rattus norvegicus	ENSRNOG00000008707	716
	Canis familiaris	ENSCAFG00000017456	752
	Monodelphis domestica	ENSMODG00000014755	725
LPO	*Homo sapiens*	ENSG00000167419	713
	Pan troglodytes	ENSPTRG00000009448	712
	Macaca mulatta	ENSMMUG00000002264	716
	Mus musculus	ENSMUSG00000009356	711
	Rattus norvegicus	ENSRNOG00000008422	710
	Canis familiaris	ENSCAFG00000024533	719
	Bos taurus	ENSBTAG00000012780	713
	Monodelphis domestica	ENSMODG00000014744	719
TPO	*Homo sapiens*	ENSG00000115705	934
	Pan troglodytes	ENSPTRG00000011610	857
	Macaca mulatta	ENSMMUG00000009662	839
	Mus musculus	ENSMUSG00000020673	915
	Rattus norvegicus	ENSRNOG00000004646	915
	Canis familiaris	ENSCAFG00000003217	932
	Bos taurus	ENSBTAG00000002567	869
	Monodelphis domestica	ENSMODG00000014296	872
PXDN	*Pan troglodytes*	5828 (PtroPxd01)*	1463
	Gallus gallus	4049 (GgaPxd01)*	1447

Note: * – Assigned entry ID and Name in the PeroxiBase database
The common names for the genomes used are; *Homo sapiens*: human, *Pan troglodytes*: chimp, *Macaca mulatta*: macaque, *Mus musculus*: mouse, *Rattus norvegicus*: rat, *Canis familiaris*: dog, *Bos taurus*: cow, *Monodelphis domestica*: opossum, *Gallus gallus*: chicken. aa: amino acid.

to generate 1000 bootstrap replicates. (The Bayesian tree reconstruction methods were applied to the MHP dataset only).

The resulting phylogenies from both analyses (MrBayes and MultiPhyl) were then analysed for signatures of LBA. The rate of evolution at each site in the alignment was placed into one of 8 categories, 8 being the most rapidly evolving and 1 being the most conserved, using the maximum likelihood approach implemented in TreePuzzle 5.1 [61]. Sites were progressively removed from the protein MSA according to their evolutionary rate and the resultant trees were analysed for changes in topology.

Nine separate site-stripped alignments were constructed by successive removal of the most rapidly evolving sites [29]. The aforementioned Bayesian method was used to infer phylogenetic relationships for each of the nine alignments generated. The ML phylogeny was also estimated for each of the site-stripped alignments from the model of best-fit following hierarchical likelihood ratio tests (hLRTs) of alternative models implemented in MultiPhyl [34].

Nodal Distance Analysis

The pruned nodal distance method implemented in TOPD/FMTS v3.3 [38] was used to calculate the distance between each of the site-stripped trees and the ideal tree.

The ideal tree was generated by pruning the resolved mammalian phylogeny [33] to represent those taxa present. A distance matrix is calculated for both the site-stripped phylogeny and the ideal phylogeny by counting the number of nodes that separate every taxon from every other taxon on the tree. Using the root means squared deviation (RMSD) implemented in the TOPD/FMTS v3.3 [38] software package, the RMSD between the site-stripped phylogeny matrix and the ideal phylogeny matrix is calculated. A RMSD value of zero indicates that the two trees being compared are identical.

Gene Tree – Species Tree Reconciliation

Following nodal distance analysis, the gene phylogeny with the lowest RMSD value (for the MHP sequences alone), and the species tree were examined for gene duplication and loss events using the default settings for gene tree – species tree reconciliation implemented in GeneTree 1.3.0 [62].

Selective Pressure Analysis

Analysis of variation in selective pressure following gene duplication in the MHPs was carried out using codon substitution models implemented in PAML 3.15 [39]. Both site-specific and branch-site specific models were applied. The models used for this analysis allow for heterogeneous nonsynonymous-to-synonymous rate ratios ($\omega = Dn/Ds$) across sites and amongst branches/lineages.

An ω-value > 1 indicates positive selection, ω < 1, purifying selection and neutral evolution when ω = 1. The statistically significant model for the data was selected using a series of LRTs to compare models and their more parameter rich extensions. Tests of significance were carried out using χ^2 tests of significance, the comparisons performed were; M0 (one ratio) with M3(k = 2)(discrete), M1(neutral) with M2(selection), M3(k = 2) with M3(k = 3) discrete models, M7 (beta) with M8 (beta & omega > 1), M8 (beta & omega > 1) with the null hypothesis M8a (beta & omega = 1), M1 with model A (branch-site) and finally M3(k = 2) with model B (branch-site). The models and approach taken here have been described previously [39,63].

The probability (PP) of a specific amino acid site belonging to the positively selected category is estimated using the empirical Bayes method for each superfamily individually [40,64,65].

Functional Divergence analysis

Using the MHP gene phylogeny with the lowest RMSD value, each of the four MHPs were selected as independent clusters. Using the MHP protein MSA and this MHP gene phylogeny, statistical analysis implemented in the software DIVERGE v 1.04 [66,46], was used to estimate the coefficient of functional divergence (theta ML or θ) for all pairs of clusters. The following are the clusters used in the analysis are taken from the resolved phylogeny (from Figure 3a) (1) MPO Cluster, (2) EPO Cluster, (3) LPO Cluster, and (4) TPO Cluster.

3D Modeling and In Silico Mutational Analysis

Homology modeling was performed using the human representative sequence for the MPO family and the first approach mode implemented by the homology-modeling server, SWISS-MODEL [48]. The structure was modeled using the crystal structure of bromide-bound human MPO isoform C (PDB accession code 1d2vC). The positively selected sites identified from the PAML 3.15 (Yang 1997) analysis were highlighted (in gold) on the 3D structure generated using DeepView v3.7 [47]. The conserved proximal heme ligand (His 502) was also highlighted (in blue) on the 3D model. *In silico* mutational analysis on these positively sites was carried out and their subsequent affect on hydrogen bonding was assessed using DeepView v3.7 [47].

Abbreviations

AD: Alzheimer's disease, AIC: Akaike Information Criterion, BEB: Bayes Empirical Bayes, DDC: Duplication-Degeneration-Complementation, Dn: nonsynonymous substitutions per nonsynonymous site, Ds: synonymous substitutions per synonymous site, EPO: Eosinophil peroxidase, hLRT: hierarchical Likelihood Ratio Test, JTT: Jones, Taylor and Thornton, LBA: Long Branch Attraction, LPO: Lactoperoxidase, LRT: Likelihood Ratio Test, MHP: Mammalian Heme Peroxidase, ML: Maximum Likelihood, MPO: Myeloperoxidase, MRCA: Most Recent Common Ancestor, MSA: Multiple Sequence Alignment, NEB: Naïve Empirical Bayes, NNI: Nearest Neighbour Interchange, PDB: Protein Data Bank, PP: Posterior Probability, PXDN: Peroxidasin, RMSD: Root Mean Squared Deviation, TIOD: Total Iodide Organification Defect, TPO: Thyroid peroxidase.

Authors' contributions

NBL carried out the alignment construction, phylogenetic and selection analysis and participated in drafting the manuscript. MJO'C designed and co-ordinated the study, and was involved in phylogeny reconstruction, selection analysis and statistical analysis, data quality control and conceived of the study. BO'C and CÓ'F were involved in the co-ordination of the study, participated in data management, contributed to the biochemical interpretation of the data and helped to draft the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

The resultant site stripped multiple sequence alignment of MHP sequences. *This figure depicts the multiple sequence alignment that was selected following RMSD analysis. This alignment has sites of rate category 8, 7, and 6 removed.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-101-S1.pdf]

Additional file 2

RMSD nodal distance between each site-stripped MHP phylogeny and the ideal phylogeny. *This table summarizes the results of the statistical comparison (RMSD) of the ideal phylogeny with each site stripped phylogeny. Values closer to zero are closer to complete agreement, the alignment with site categories 8 through to 6 removed, is the phylogeny closest to ideal.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-101-S2.doc]

Additional file 3

Parameter estimates and likelihood scores of one ratio and site-specific models. *The data presented in this table are the results of ML analysis of site specific evolutionary models applied to the MHP alignment. The name of the model is given in column 1, the number of parameters estimated is given in column 2, the Log likelihood value in column 3, and the parameter estimates in column 4 and 5.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-101-S3.doc]

Additional file 4

Parameter estimates and likelihood scores for branch-site models: MPO and EPO clades. *This table summarizes the results of ML analysis on the MHP data, using branch specific models of evolution. The MPO and EPO clades are treated as foreground lineages independently and all other peroxidase clades as background. The LRTs are performed between model A and M1 and model B and M3K2 from Additional file 3.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-101-S4.doc]

Additional file 5

Parameter estimates and likelihood scores for branch-site models: LPO and TPO clades. *This table summarizes the results of ML analysis on the MHP data, using branch specific models of evolution. The LPO and TPO clades are treated as foreground lineages independently and all other peroxidase clades as background. The LRTs are performed between model A and M1 and model B and M3K2 from Additional file 3.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-101-S5.doc]

Acknowledgements

We would like to thank the Irish Research Council for Science, Engineering and Technology (Embark Initiative Postgraduate Scholarship to NBL) for financial support. We would like to thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for processor time and technical support for both phylogeny reconstruction and selection analysis. We would like to acknowledge Dr James McInerney's research group at the Bioinformatics Laboratory, NUI Maynooth for the use of their computational facilities. We would like to thank Dr Christopher Creevey, European Molecular Biology Laboratory, Heidelberg, Germany for generously supplying us with the necessary computer code for conducting our site-stripping analysis.

References

1. Passardi F, Bakalovic N, Teixeira FK, Margis-Pinheiro M, Penel C, Dunand C: **Prokaryotic origins of the non-animal peroxidase superfamily and organelle-mediated transmission to eukaryotes.** *Genomics* 2007, **89(5):**567-579.
2. Dunford HB: **Heme Peroxidases.** New York , John Wiley and Sons Inc; 1999.
3. Furtmuller PG, Zederbauer M, Jantschko W, Helm J, Bogner M, Jakopitsch C, Obinger C: **Active site structure and catalytic mechanisms of human peroxidases.** *Archives of biochemistry and biophysics* 2006, **445(2):**199-213.
4. Clark RA: **Peroxidases: A historical overview of Milestones in Research on Myeloperoxidase.** In *The peroxidase multigene family of enzymes: biochemical basis and clinical applications* Edited by: Petrides, P.E., Nauseef WM. Heidleberg , Springer-Verlag; 2000:1-10.
5. Sanz ML, Parra A, Prieto I, Dieguez I, Oehling AK: **Serum eosinophil peroxidase (EPO) levels in asthmatic patients.** *Allergy* 1997, **52(4):**417-422.
6. Reynolds WF, Hiltunen M, Pirskanen M, Mannermaa A, Helisalmi S, Lehtovirta M, Alafuzoff I, Soininen H: **MPO and APOEepsilon4 polymorphisms interact to increase risk for AD in Finnish males.** *Neurology* 2000, **55(9):**1284-1290.
7. Lau D, Baldus S: **Myeloperoxidase and its contributory role in inflammatory vascular disease.** *Pharmacology & therapeutics* 2006, **111(1):**16-26.
8. Sakamaki K, Kanda N, Ueda T, Aikawa E, Nagata S: **The eosinophil peroxidase gene forms a cluster with the genes for myeloperoxidase and lactoperoxidase on human chromosome 17.** *Cytogenetics and cell genetics* 2000, **88(3–4):**246-248.
9. Sakamaki K, Ueda T, Nagata S: **The evolutionary conservation of the mammalian peroxidase genes.** *Cytogenetic and genome research* 2002, **98(1):**93-95.
10. Ohno S: **Evolution by gene duplication.** New York , Springer-Verlag; 1970.
11. Nowak MA, Boerlijst MC, Cooke J, Smith JM: **Evolution of genetic redundancy.** *Nature* 1997, **388(6638):**167-171.
12. Chung WY, Albert R, Albert I, Nekrutenko A, Makova KD: **Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network.** *BMC bioinformatics* 2006, **7:**46.
13. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 1999, **151(4):**1531-1545.
14. Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH: **Role of positive selection in the retention of duplicate genes in mammalian genomes.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103(7):**2232-2236.
15. Cheng JD, Ryseck RP, Attar RM, Dambach D, Bravo R: **Functional redundancy of the nuclear factor kappa B inhibitors I kappa B alpha and I kappa B beta.** *The Journal of experimental medicine* 1998, **188(6):**1055-1062.
16. Hughes T, Liberles DA: **The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation.** *Journal of molecular evolution* 2007, **65(5):**574-588.
17. Kimura S, Ikeda-Saito M: **Human myeloperoxidase and thyroid peroxidase, two enzymes with separate and distinct physiological functions, are evolutionarily related members of the same gene family.** *Proteins* 1988, **3(2):**113-120.
18. Hughes AL: **Adaptive evolution of genes and genomes.** 1st edition. New York , Oxford University Press; 1999:270.
19. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154(1):**459-473.
20. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA: **Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms.** *Journal of experimental zoology Part B* 2007, **308(1):**58-73.

21. Klebanoff SJ: **Myeloperoxidase: contribution to the microbicidal activity of intact leukocytes.** *Science (New York, NY* 1970, **169(950):**1095-1097.
22. Klebanoff SJ: **Myeloperoxidase.** *Proceedings of the Association of American Physicians* 1999, **111(5):**383-389.
23. Wang J, Slungaard A: **Role of eosinophil peroxidase in host defense and disease pathology.** *Archives of biochemistry and biophysics* 2006, **445(2):**256-260.
24. Ruf J, Carayon P: **Structural and functional aspects of thyroid peroxidase.** *Archives of biochemistry and biophysics* 2006, **445(2):**269-277.
25. Zederbauer M, Furtmuller PG, Brogioni S, Jakopitsch C, Smulevich G, Obinger C: **Heme to protein linkages in mammalian peroxidases: impact on spectroscopic, redox and catalytic properties.** *Natural product reports* 2007, **24(3):**571-584.
26. Moreira D, Philippe H: **Molecular phylogeny: pitfalls and progress.** *Int Microbiol* 2000, **3(1):**9-16.
27. Ohta T: **An examination of the generation-time effect on molecular evolution.** *Proceedings of the National Academy of Sciences of the United States of America* 1993, **90(22):**10676-10680.
28. Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D: **Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis.** *Molecular phylogenetics and evolution* 1996, **5(1):**182-187.
29. Philip GK, Creevey CJ, McInerney JO: **The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa.** *Molecular biology and evolution* 2005, **22(5):**1175-1184.
30. **Peroxibase** [http://peroxibase.isb-sib.ch]
31. Passardi F, Theiler G, Zamocky M, Cosio C, Rouhier N, Teixera F, Margis-Pinheiro M, Ioannidis V, Penel C, Falquet L, Dunand C: **PeroxiBase: the peroxidase database.** *Phytochemistry* 2007, **68(12):**1605-1611.
32. Levasseur A, Gouret P, Lesage-Meessen L, Asther M, Asther M, Record E, Pontarotti P: **Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family.** *BMC evolutionary biology* 2006, **6:**92.
33. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ: **Molecular phylogenetics and the origins of placental mammals.** *Nature* 2001, **409(6820):**614-618.
34. Keane TM, Naughton TJ, McInerney JO: **MultiPhyl: a high-throughput phylogenomics webserver using distributed computing.** *Nucleic acids research* 2007, **35(Web Server issue):**W33-7.
35. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics (Oxford, England)* 2003, **19(12):**1572-1574.
36. Lunter G: **Dog as an outgroup to human and mouse.** *PLoS computational biology* 2007, **3(4):**e74.
37. Brinkmann H, Philippe H: **Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies.** *Molecular biology and evolution* 1999, **16(6):**817-825.
38. Puigbo P, Garcia-Vallve S, McInerney JO: **TOPD/FMTS: a new software to compare phylogenetic trees.** *Bioinformatics (Oxford, England)* 2007, **23(12):**1556-1558.
39. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13(5):**555-556.
40. Yang Z, Nielsen R, Goldman N, Pedersen AM: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155(1):**431-449.
41. Andersson E, Hellman L, Gullberg U, Olsson I: **The role of the propeptide for processing and sorting of human myeloperoxidase.** *The Journal of biological chemistry* 1998, **273(8):**4747-4753.
42. Kizaki M, Miller CW, Selsted ME, Koeffler HP: **Myeloperoxidase (MPO) gene mutation in hereditary MPO deficiency.** *Blood* 1994, **83(7):**1935-1940.
43. Rietsch A, Beckwith J: **The genetics of disulfide bond metabolism.** *Annual review of genetics* 1998, **32:**163-184.
44. Ferrand M, Le Fourn V, Franc JL: **Increasing diversity of human thyroperoxidase generated by alternative splicing. Characterized by molecular cloning of new transcripts with single- and multispliced mRNAs.** *The Journal of biological chemistry* 2003, **278(6):**3793-3800.
45. Tajima T, Tsubaki J, Fujieda K: **Two novel mutations in the thyroid peroxidase gene with goitrous hypothyroidism.** *Endocrine journal* 2005, **52(5):**643-645.
46. Gu X, Vander Velden K: **DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family.** *Bioinformatics (Oxford, England)* 2002, **18(3):**500-501.
47. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18(15):**2714-2723.
48. Arnold K, Bordoli L, Kopp J, Schwede T: **The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling.** *Bioinformatics (Oxford, England)* 2006, **22(2):**195-201.
49. Martin AC, Facchiano AM, Cuff AL, Hernandez-Boussard T, Olivier M, Hainaut P, Thornton JM: **Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein.** *Human mutation* 2002, **19(2):**149-164.
50. Li WH, Tanimura M, Sharp PM: **An evaluation of the molecular clock hypothesis using mammalian DNA sequences.** *Journal of molecular evolution* 1987, **25(4):**330-342.
51. Page RD, Cotton JA: **Vertebrate phylogenomics: reconciled trees and gene duplications.** *Pacific Symposium on Biocomputing* 2002:536-547.
52. Grebski E, Peterson C, Medici TC: **Effect of physical and chemical methods of homogenization on inflammatory mediators in sputum of asthma patients.** *Chest* 2001, **119(5):**1521-1525.
53. Ambrugger P, Stoeva I, Biebermann H, Torresani T, Leitner C, Grutters A: **Novel mutations of the thyroid peroxidase gene in patients with permanent congenital hypothyroidism.** *European journal of endocrinology / European Federation of Endocrine Societies* 2001, **145(1):**19-24.
54. Bikker H, Baas F, De Vijlder JJ: **Molecular analysis of mutated thyroid peroxidase detected in patients with total iodide organification defects.** *The Journal of clinical endocrinology and metabolism* 1997, **82(2):**649-653.
55. Rodrigues C, Jorge P, Soares JP, Santos I, Salomao R, Madeira M, Osorio RV, Santos R: **Mutation screening of the thyroid peroxidase gene in a cohort of 55 Portuguese patients with congenital hypothyroidism.** *European journal of endocrinology / European Federation of Endocrine Societies* 2005, **152(2):**193-198.
56. **Ensembl** [http://www.ensembl.org]
57. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC bioinformatics* 2004, **5:**113.
58. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic acids research* 1994, **22(22):**4673-4680.
59. Rambaut A: **SE-AL Sequence alignment editor.** 2.0a11 edition. 1996 [http://evolve.zoo.ox.ac.uk/]. Oxford
60. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8(3):**275-282.
61. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics (Oxford, England)* 2002, **18(3):**502-504.
62. Page RD: **GeneTree: comparing gene and species phylogenies using reconciled trees.** *Bioinformatics (Oxford, England)* 1998, **14(9):**819-820.
63. O'Connell MJ, McInerney JO: **Gamma chain receptor interleukins: evidence for positive selection driving the evolution of cell-to-cell communicators in the mammalian immune system.** *Journal of molecular evolution* 2005, **61(5):**608-619.
64. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148(3):**929-936.
65. Yang Z, Wong WS, Nielsen R: **Bayes empirical bayes inference of amino acid sites under positive selection.** *Molecular biology and evolution* 2005, **22(4):**1107-1118.
66. Gu X: **Statistical methods for testing functional divergence after gene duplication.** *Molecular biology and evolution* 1999, **16(12):**1664-1674.