EGIA - Evolutionary optimisation of Gene regulatory networks, an Integrative Approach

Alina Sîrbu, Martin Crane and Heather J. Ruskin

Abstract Quantitative modelling of gene regulatory networks (GRNs) is still limited by data issues such as noise and the restricted length of available time series, creating an under-determination problem. However, large amounts of other types of biological data and knowledge are available, such as knockout experiments, annotations and so on, and it has been postulated that integration of these can improve model quality. However, integration has not been fully explored to date. Here, we present a novel integrative framework for different types of data that aims to enhance model inference. This is based on evolutionary computation and uses different types of knowledge to introduce a novel *customised initialisation and mutation operator* and *complex evaluation criteria*, used to distinguish between candidate models. Specifically, the algorithm uses information from (i) knockout experiments, (ii) annotations of transcription factors, (iii) binding site motifs (expressed as position weight matrices) and (iv) DNA sequence of gene promoters, to drive the algorithm towards more plausible network structures. Further, the evaluation basis is also extended to include structure information included in these additional data. This framework is applied to both synthetic and real gene expression data. Models obtained by data integration display both quantitative and qualitative improvement.

Alina Sîrbu
Institute for Scientific Interchange Foundation, Turin, Italy ; Center for Scientific Computing and Complex Systems Modelling
School of Computing, Dublin City University, Dublin, Ireland e-mail: alina.sirbu@isi.it

Martin Crane
Center for Scientific Computing and Complex Systems Modelling
School of Computing, Dublin City University, Dublin, Ireland

Heather J. Ruskin
Center for Scientific Computing and Complex Systems Modelling
School of Computing, Dublin City University, Dublin, Ireland

1 Introduction

Gene regulatory network reverse engineering is an important aim of Systems Biology [7], as models obtained can be used for analysis and simulation in contexts often difficult to realise in laboratory experiments. Approaches using mathematical modelling, ranging from qualitative to quantitative, have been applied to discovery of GRNs from gene expression data [10]. However, the size of GRNs and the nature of the data (high dimensional, noisy, insufficient for analysis of dynamics), limit robustness when mimicking natural behaviour. This is particularly true for *quantitative* models, which aim to simulate very detailed patterns of expression, increasing the number of parameters to be inferred. However, such models can provide extremely useful insight on the gene expression process, where improvement of reverse engineering techniques is an ongoing aim of Systems Biology [16].

Given the challenges posed by available gene expression data and poor model robustness to date, a new direction is integration of several data types, [16], and these reports have started to appear, mostly for coarse-grained analysis [8]. These integrate expression data with other types of measurements, such as binding affinities or protein interactions, to better discriminate between candidate models, but usually are limited, (i.e. use only one additional data type, besides time-course data). However, several such data-types are available, and the hypothesis is that combining all of these, can further increase modelling power. Recently, *Drosophila Melanogaster* datasets have been integrated, but again for qualitative analysis only [3]. Here, a novel inferential framework for *quantitative* models, based on Evolutionary Computation (EC), is presented (EGIA - Evolutionary optimisation of GRNs - an Integrative Approach). Although other methods are also possible, the EC approach has been selected as it provides increased flexibility, implicit parallelism and has proved to be a suitable search method for underdetermined problems, noisy data and large search spaces [1]. The hypothesis tested is that integration of diverse large-scale biological data improves qualitative and quantitative performance of models inferred.

The strength of the newly-introduced platform is the number of data types to be combined and the flexibility of integration. The novel customisation of different stages of the Evolutionary Algorithm permits more knowledgeable exploration of the search space and more informative evaluation criteria, based on the data available. This is crucial for improving the performance of the models inferred, both quantitatively and qualitatively. Furthermore, a general methodology for GRN inference from multiple data types is developed. This includes an *error structure analysis* to identify the stage of the algorithm at which each data type should be integrated.

2 Methods

2.1 Data

Both synthetic and real datasets are used to assess algorithm performance. Synthetic networks are from the DREAM4[12] competition. This is a research community competition where data from known GRNs are published and researchers have the

task of reverse engineering the original networks. These networks are carefully generated so as to resemble real GRNs. The data used here, generated by networks of 10 and 100 genes respectively, contain both time-series measurements and knockout experiments. The set of known interactions are used for qualitative evaluation, and MSE for dual-knockout experiments for quantitative.

For real data, a sub-network of 27 genes involved in *Drosophila melanogaster* embryo development is analysed. A single-channel (SC) microarray dataset [21], is used for training, while a dual-channel (DC) dataset [11] is used for quantitative evaluation. Cross-platform normalisation (namely XPN, [17]) has been performed prior to model inference. For qualitative evaluation, 16 interactions from the Drosophila Interactions Database (DROID) [13], version 2010_10, are considered gold-standard. Additional data types are also integrated: (i) knockout experiments for 8 genes, which were used to compute log-ratios against wild-type experiments [11, 5, 20, 4, 6], (ii) pair-wise correlation between gene expression patterns, (iii) Gene Ontology (GO) [14] annotations, which assign the function of *transcriptional regulation* to 17 of these genes and (iv) binding site affinities for 11 transcription factors (computed using known cis-regulatory modules and position specific weight matrices - PSWMs [15, 2]).

Algorithm performance is evaluated both quantitatively and qualitatively. *Qualitative* evaluation analyses GRN *topology*, to assess whether known interactions between gene pairs are retrieved by the algorithm. This means that the known adjacency matrix of the network is compared to the one retrieved by our algorithm. Specifically, the AUROC (Area Under the ROC Curve) and AUPR (Area Under the Precision-Recall Curve) are computed, measures used also in the DREAM4 competition. Given that the algorithm is stochastic in nature, predictions of interactions have been performed by using multiple models obtained in different runs, and employing a voting procedure for possible interactions. In this way, an interaction that appears in more models is considered to be more plausible. The ranking of possible interactions is used for AUROC/AUPR computation. *Quantitative* evaluation assesses whether the inferred models are able to predict the real-valued expression levels seen in the data. This is performed by simulating a set of *test* data, *not used for model inference*, and by computing average MSE (Mean Squared Error) values over multiple runs.

2.2 Algorithm

EGIA seeks to exploit several types of data related to the process of gene expression, which contribute at different stages of the evolutionary algorithm. The framework is based on a previously introduced inferential algorithm, [9]. This algorithm has been shown to be among the most scalable and least sensitive to noise of several methods from the literature [18]. Based on this, we have chosen to extend it further for data integration, by introducing *novel mutation, initialisation and evaluation operators*.

2.2.1 The basic algorithm

In [9], a neural-genetic hybrid approach to GRN inference was introduced. This models the GRN as a single-layered Artificial neural Network (ANN), consisting of one neural unit per gene. Each unit i takes as input the expression values of the regulators of gene g_i (i.e. g_j) at time point t and computes the expression level for gene g_i at time $t+1$, using the input weights w_{ij} and the logistic function $S(x) = \frac{1}{1+e^{-x}}$ for activation:

$$g_i(t+1) = S\left(\sum_j w_{ij} g_j(t) + b_i - d_i g_i(t)\right) \tag{1}$$

where b_i accounts for external input, while d_i represents the degradation rate.

The basic algorithm divides optimisation into two phases: *structure* and *parameter* search. The first involves optimising network topology, i.e. the set of regulators for each gene. This is implemented as a Genetic Algorithm, where each individual encodes a candidate structure, as a subset of the possible regulators for the current gene. Each candidate structure is assigned a fitness value during the parameter search phase, which employs Gradient Descent to optimise the input weights for the neural unit for the current gene. The final error obtained is considered the fitness of the candidate structure. A divide-and-conquer approach is used to optimise parameters for each gene at a time, i.e. training small networks with one neural unit, independently of the other units.

2.2.2 Algorithmic schema extension

The basic algorithm [9] optimises parameters for each gene separately, in a divide-and-conquer manner. This approach reduces dimensionality of the system for each optimisation run. However, the model obtained by directly combining sub-models may not be able to correctly simulate the whole system, as separate optimisation disregards the feed-back from the full gene set. In consequence, we have added a second optimisation stage, which combines single-gene models and performs a fine-tuning of complete-network parameters, using the same structure and parameter optimisation.

One way of obtaining models that are robust to noise involves creating noisy replicates from the available data [22]. This simulates technical replicates, and results in multiple time series to be used during inference. Here, a larger set of time-series has been derived from available data through addition of random Gaussian noise. This has been performed during the parameter optimisation phase, for ANN training.

2.2.3 Custom initialisation and mutation

The basic algorithm achieves an initial population of candidate structures by randomly selecting possible transcription factors for a specific gene. Similarly, mutation is performed by replacing one of the regulators with a randomly chosen gene.

However, many data types provide indications on which interactions between genes are most likely. For example, binding site affinities can indicate what transcription factors can bind to a specific gene promoter. This type of information is very valuable, and can be used to explore the search space in a more knowledgeable manner. For this, we have developed a *customised initialisation and mutation* procedure, which uses likelihood assignment for gene regulation. This results, for each gene g, in a non-uniform probability mass function, which describes which of the genes in the network are more likely to be regulators of gene g. When performing mutation or initialisation, this function is used to select a candidate regulator for gene g. This is similar to *Wheel of Fortune (WOF)* selection [1], (also known as the *roulette wheel*), so will be addressed henceforth as WOF mutation and initialisation.

In order to build the probability mass function for each gene g, the strategy is to assign segments on the WOF to each gene in the network, if there is any indication in the data of a possible effect of that gene on the current gene g. This number of segments has to be defined by the user based on the reliability of the data used. In the following we provide the values used in our experiments, empirically determined through multiple applications of the algorithm. Of course, these values can be changed to produce a higher or lower effect on the resulting WOF. Several different types of data can be used for this, as follows.

Correlation patterns Althought dependences between genes can be non-linear, a good correspondence between linear gene expression correlation-based networks and GRNs has been previously identified, [23]. In consequence, we have used Pearson correlation between time series data of gene pairs, to enhance solution space exploration. Based on absolute values of the correlation to gene g, each gene i is assigned segments on the WOF:

$$CORR_{gi} = \begin{cases} 0 & \text{if } |r_{gi}| < \text{1st decile} \\ 1 & \text{if 1st decile} < |r_{gi}| < \text{3rd decile} \\ 4 & \text{if 3rd decile} < |r_{gi}| < \text{7th decile} \\ 6 & \text{otherwise} \end{cases} \tag{2}$$

where r_{gi} is the Pearson coefficient between genes i and g. The deciles are based on all correlation values obtained. In this way, genes that show high correlation with the current gene will be more likely to be selected as possible regulators.

Knockout(KO) experiments Gene expression data from KO experiments can also be used to enhance the search for network models. Absolute values of log-ratios between wild-type and knockout samples can be used to allocate segments on the WOF to those genes that display a large effect on other genes. The number of segments (KO_{gi}) allocated for each gene i on the WOF of gene g depends on the magnitude of the log-ratio:

$$KO_{gi} = \begin{cases} 0 & \text{if } |\text{log-ratio}_{gi}| < 0.2 \\ 1 & \text{if } 0.2 < |\text{log-ratio}_{gi}| < 0.5 \\ 4 & \text{if } 0.5 < |\text{log-ratio}_{gi}| < 0.8 \\ 6 & \text{if } 0.8 < |\text{log-ratio}_{gi}| < 1.1 \\ 8 & \text{otherwise} \end{cases} \tag{3}$$

Gene Ontology (GO) annotations The GO database contains annotations of which gene products have been observed to have a specific function, and annotations of transcriptional regulator activity can be included in the EGIA framework. These genes will be allocated additional segments (4 in our experiments) on all the wheels of fortune of the genes in the network. In this way, known transcription factors become more likely to be selected as regulators:

$$ANNOT_{gi} = \begin{cases} 0 & \text{if gene } i \text{ is not annotated as TF} \\ 4 & \text{otherwise} \end{cases} \tag{4}$$

Binding site affinities Binding site (BS) affinities can be integrated in a similar manner. To compute the affinity between a regulator and a gene, the position specofic weight matrix (PSWM) associated with the regulator is required, as well as promoter sequences for the gene. Using these two pieces of information, BS affinity values for each TF i and target gene g are retrieved. For each regulator i, the average ($\overline{A}$) and maximum affinity (A_{max}), over all target genes g, is computed, and segments on the WOF are allocated as follows:

$$BS_{gi} = \begin{cases} 0 & \text{if } A_{gi} < \overline{A} \\ 6 & \text{if } \overline{A} < A_{gi} < \overline{A} + \frac{\overline{A} - A_{max}}{2} \\ 8 & \text{otherwise} \end{cases} \tag{5}$$

where A_{gi} represents the affinity of gene i for binding to a promoter of gene g.

Once all the segments, corresponding to the different type of data, are allocated for all possible regulators, these are summed (Equation 6) and the segment distribution is normalised to represent a probability mass function (Equation 7).

$$WOF_{gi} = CORR_{gi} + KO_{gi} + BS_{gi} + ANNOT_{gi} \tag{6}$$

$$f_g(i) = \frac{WOF_{gi}}{\sum_i WOF_{gi}} \tag{7}$$

This probability mass function defines the probability that a gene i will be selected as regulator for gene g during mutation and initialisation. Each target gene g is associated with such a probability mass function. All data types mentioned can be integrated or omitted, depending on availability. When no additional data are available, the WOF mutation and initialisation are equivalent to the random assignment from the basic algorithm.

2.2.4 Extending evaluation

The original algorithm uses a fitness function based on the RSS between data and simulation. This has been extended to include also the correlation between simulation and gene patterns [19]. However, this only considers time-series data for evaluation. Using additional data during model evaluation, which might provide information on possible structure, is one way of addressing the noise and underdetermination problem, inherent in time-series data. This changes the fitness land-

scape, so that models which have a plausible topology as well as ability to simulate the time-series data, correspond to better fitness.

The WOF mechanism presented in Section 2.2.3 can be thus also used for model evaluation, by computing an average of all probabilities assigned to the model interactions by the WOF. This, used in combination with the previous fitness function discussed [19], enables construction of a fitness landscape that helps the optimisation algorithm find more plausible structures, as well as models that can simulate continuous behaviour. The final fitness function to be minimised is:

$$F = \frac{1}{2}\sum_i (o_i - t_i)^2 - cP - w\frac{1}{n}\sum_{(i,j)\in INT} f_j(i) \tag{8}$$

where the first term on the right hand side represents the squared error typical for ANN backpropagation (o_i is the expression level simulated by the model, while t_i is that observed in the data), the second the correlation term from [19], while the last term is an average, over all pair-wise interactions present in the model, of the probabilities obtained by the WOF mechanism. INT is the set of interactions predicted by the model ((i, j) is an inferred regulatory effect of i on j), while $f_j(i)$ represents the fraction of the WOF allocated to that interaction (Equation 7). This term is weighted by w, a parameter which needs to be provided by the user. This evaluation criterion is used both at the single-gene and complete-model optimisation stage.

3 Results

The customised evolutionary operators have been implemented using all data types available and models obtained compared to the original algorithm. In order to identify which type of data is more useful, different variants of WOF and evaluation have also been employed, by eliminating one data type at a time.

3.1 Performance on synthetic networks

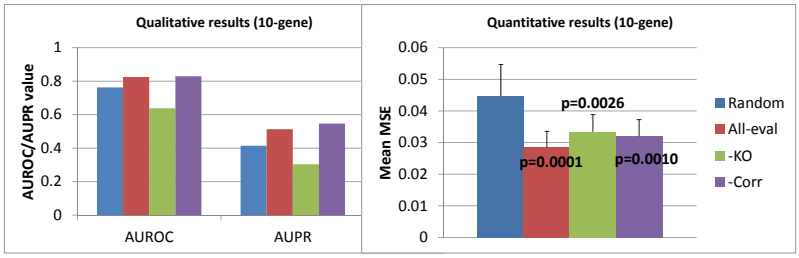

Fig. 1 Performance of WOF and extended evaluation for the 10-gene synthetic dataset.

For the synthetic datasets (DREAM networks), only correlation patterns and log-ratios for knockout experiments are available, so three versions of the algorithm were compared to the basic one (Random). These three variants are denoted by *All-eval* (including all data available in WOF mutation, initialisation and evaluation), *-KO* (all data excluding knockout experiments) and *-Corr* (all data excluding correlation patterns).

Figure 1 displays AUROC and AUPR values obtained after 10 runs of each algorithm on the 10-gene synthetic network. It also includes average MSE over 10 runs for dual knockout simulations, and corresponding p-values of differences observed (compared to the basic algorithm - *Random*). As the figures show, extending the evaluation criterion appears to produce both qualitative and quantitative improvement when compared to the basic algorithm. The set of predicted interactions is slightly improved when knockout experiments only are used (*-Corr*), but quantitative behaviour is best (lowest MSE values) when both data types are integrated. However, when knockout experiments are excluded, AUROC/AUPR values decrease significantly. This suggests that knockout data are very important for extracting direct interactions.

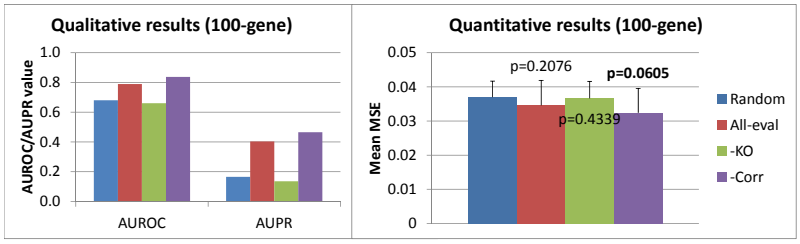

Fig. 2 WOF and extended evaluation for the 100-gene synthetic dataset.

Similarly, for the 100-gene network, qualitative and quantitative results are displayed in Figure 2. Introducing the enhanced evaluation criterion markedly increases the number of correct interactions discovered, as shown by the AUROC and AUPR values. The best results are obtained after excluding correlation patterns from the data types used, indicating again that these are not particularly useful in this context, (as found also for the 10-gene network). On the other hand, if knockout experiments are excluded, AUROC/AUPR values decrease significantly, showing that these data are very important in predicting a good set of interactions. From the quantitative point of view, the novel evaluation criterion yields models with low MSE in dual knockout simulations, (minimum values under 0.025), with best results obtained for exclusion of correlation patterns. However, although minimum and average MSE are lower compared to the basic algorithm, the overall quantitative results from multiple experiments are only statistically significant at the 10% level (*-Corr*).

We have compared these results to those obtained by the participants in the DREAM4 competition, on the same networks used in this analysis. The top three teams, which submitted *quantitative* and *qualitative* results for *both network sizes*,

have been selected for comparison. For these, AUROC/AUPR and MSE values are given in Table 1, with best performances outlined in bold font. EGIA has obtained the *best predicted interactions for the large scale network*, while for the small scale it scored 3rd. This indicates that our method is more scalable compared to the others. From the quantitative simulation point of view, EGIA has obtained models with lower MSE than the other methods on dual knockouts for both network sizes although, on average, behaviour is comparable to other methods. Nevertheless, given the good qualitative results, we conclude that this framework has something to contribute for extracting models with correct interactions, while it can also simulate unseen behaviour.

Table 1 Comparison of EGIA with DREAM4 results. For the dual knockout MSE values of EGIA, both the minimum and the average values obtained in repeated runs are provided.

	10-gene $\sqrt{AUROC} * \sqrt{AUPR}$	10-gene dual-KO MSE	100-gene $\sqrt{AUROC} * \sqrt{AUPR}$	100-gene dual-KO MSE
EGIA	0.6735	**0.019**/0.028	**0.624**	**0.0229**/0.0324
Team 548	0.654	0.038	0.544	0.0349
Team 532	**0.733**	0.020	0.505	0.0303
Team 498	0.702	0.029	0.28	0.0327

3.2 Performance on the Drosophila network

For the real dataset, five variants of the algorithm have been analysed: *All-eval* (evaluation and WOF operators using all data available), *-Corr* (all data excluding correlation patterns), *-KO* (excluding knockout experiments), *-BS* (excluding binding site affinities), *-Annot* (excluding GO annotations), enabling assessment of the error structure in these data and how this influences the models obtained. Figure 3 displays AUROC and AUPR values for the five algorithm variants. These indicate that integrating all types of data yields the best prediction for interactions. The largest effect is from the binding site affinity data. However, all data types seem to contribute, unlike the synthetic data where correlation patterns disimproved performance compared to the basic algorithm.

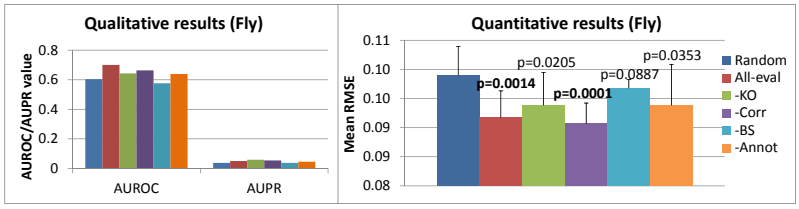

Fig. 3 Performance of WOF and extended evaluation for the 27-gene real dataset.

Quantitative evaluation was performed again by computing the RMSE with the test dataset (DC), and Figure 3 also shows average results obtained by each of the

algorithm variants in 10 runs, with p-values of observed differences from the basic algorithm. Our algorithm improves quantitative behaviour, with RMSE values significantly lower than the basic algorithm (at the 1% level for *All-eval* and *-Corr*, and the 5% level for *-KO* and *-Annot*). This improvement means that models not only contain more valid interactions, but also simulate test data better, i.e. improvement in both qualitative and quantitative performance. The error structure analysis also indicates that correlation patterns are once again not particularly useful for improving quantitative performance, while binding site affinities seem to be crucial.

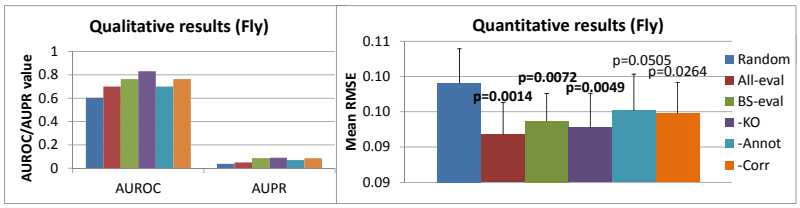

Fig. 4 WOF and binding site extended evaluation for the 27-gene real dataset.

While WOF is a *weak* integration method, as it drives the algorithm only towards promising areas of the search space, without forcing it to choose one model or another, extended evaluation is a *strong* integration criterion, having the final say in which model is better. So, while the WOF operators can be resilient to some level of noise in the data, the evaluation criterion must include more specific data types. Given the results from the error structure analysis for the real dataset, correlation patterns, knockout experiments and GO annotation are more suitable for WOF alone, as they provide *guideline* information only on potential interactions. Binding site affinities are, however, suitable for formal model evaluation, as they have proved to be crucial for obtaining good quantitative performance (Figure 3). For the rest of this section, therefore, we present a similar analysis for different algorithm variants employing only binding site affinities in evaluation, but using various forms of WOF operators: *BS-eval* (using all data types for WOF), *-Corr* (excluding correlation patterns from WOF), *-KO* (excluding knockout experiments), *-Annot* (excluding GO annotations).

Figure 4 displays the performance for all four algorithm variants above, compared to *All-eval* (evaluation and WOF using all data types) and *Random*, the basic algorithm (no meta-data used). *BS-eval* produces models with better connections compared to *All-eval*, while RMSE on test data is maintained at a low level (*BS-eval* and *-KO* significantly different from *Random* at the 1% level).

On extending evaluation, RMSE values for training data display a slight increase, both for synthetic and real data. One explanation for this is that the *generalisation* ability of models is increased (RMSE on test data decreases), and the *over-fitting* of training data is decreased. Generally, machine learning techniques need to obtain a balance between generalisation and over-fitting, which was made possible here by the inclusion of additional data types for training.

4 Conclusion

This paper presented an analysis of data integration for GRN modelling. Two integration mechanisms have been analysed, namely *customised mutation and initialisation* (WOF) and *extended evaluation*. The *error structure* of available data has been studied, to identify which data type has larger effect on the networks analysed. WOF and extended evaluation led to both quantitative and qualitative improvement. This supports the hypothesis that optimisation with time-series data alone is not powerful enough, and that additional information from other data types is needed to aid further selection of GRN models.

The error structure analysis suggested that not all data types are useful for inference, however, and that great caution needs to be taken when integrating these. For synthetic data, knockout experiments proved to be highly important to improve predictions of regulatory interactions. For real data, binding site affinities seemed to have the largest impact. Correlation patterns, on the other hand, were of some help when integrated in WOF mutation with other data types, but had less individual importance. This might be due to the fact that correlation does not indicate only direct interaction, but also indirect effects, which can be captured by the models.

WOF proved to be a flexible integration tool, while evaluation provided additional rigour. For best results, only very reliable data should be used for the latter, while noisy data can be integrated into the former, following an error structure analysis. In our experiments, best performance on real data was found by using only binding affinities for evaluation, and all data types for WOF. This suggests that other data types can provide only general guidelines for possible structures. For instance, log ratios in knockout experiments, or correlations between gene expression patterns can sometimes be misleading, due to the existence of feedback loops, related to alternative regulatory paths or indirect interactions in the real network. The results presented here apply for the *Drosophila melanogaster* embryo development network and associated datasets available for this system. In analysing other systems, e.g. different processes or organisms, data types and quality available will vary, so performing an initial error analysis is crucial to determining the best integration strategy.

Acknowledgements We would like to thank the Irish Centre for High-End Computing for the provision of computational resources for the experiments. This work was supported by the Irish Research Council for Science, Engineering and Technology, under the EMBARK scheme and partly by the EveryAware project funded by the EC under the EU RD contract IST-265432.

References

1. Baeck, T., Fogel, D.B., Michalewicz, Z.: Evolutionary Computation 1: Basic Algorithms and Operators. Institute of Physics Publishing, Bristol and Philadelphia (2000)
2. Bergman, C.M., Carlson, J.W., Celniker, S.E.: Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. Bioinformatics **21**(8), 1747–1749 (2005)

3. modENCODE Consortium, T.: Identification of functional elements and regulatory circuits by drosophila modencode. Science (2010)
4. Elgar, S.J., Han, J., Taylor, M.V.: Mef2 activity levels differentially affect gene expression during drosophila muscle development. Proceedings of the National Academy of Sciences of the United States of America **105**(3), 918–923 (2008)
5. Estrada, B., Choe, S.E., Gisselbrecht, S.S., Michaud, S., Raj, L., Busser, B.W., Halfon, M.S., Church, G.M., Michelson, A.M.: An integrated strategy for analyzing the unique developmental programs of different myoblast subtypes. PLoS Genetics **2**(2), e16 (2006)
6. Fox, R.M., Hanlon, C.D., Andrew, D.J.: The CrebA/Creb3-like transcription factors are major and direct regulators of secretory capacity. The Journal of Cell Biology **191**(3), 479–492 (2010)
7. Heath, A., Kavraki, L.: Computational challenges in systems biology. Computer Science Review **3**(1), 1–17 (2009)
8. Huttenhower, C., Mutungu, K.T., Indik, N., Yang, W., Schroeder, M., Forman, J.J., Troyanskaya, O.G., Coller, H.A.: Detailing regulatory networks through large scale data integration. Bioinformatics **25**(24), 3267–3274 (2009)
9. Keedwell, E., Narayanan, A.: Discovering gene networks with a neural-genetic hybrid. Computational Biology and Bioinformatics, IEEE/ACM Transactions on **2**(3), 231–242 (2005)
10. Lee, W.P., Tzou, W.S.: Computational methods for discovering gene networks from expression data. Briefings in Bioinformatics **10**(4), 408–423 (2009)
11. Liu, J., Ghanim, M., Xue, L., Brown, C.D., Iossifov, I., Angeletti, C., Hua, S., Negre, N., Ludwig, M., Stricker, T., Al-Ahmadie, H.A., Tretiakova, M., Camp, R.L., Perera-Alberto, M., Rimm, D.L., Xu, T., Rzhetsky, A., White, K.P.: Analysis of *Drosophila* Segmentation Network Identifies a JNK Pathway Factor Overexpressed in Kidney Cancer. Science **323**(5918), 1218–1222 (2009)
12. Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D., Stolovitzky, G.: Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences of the United States of America **107**(14), 6286–6291 (2010)
13. Murali, T., Pacifico, S., Yu, J., Guest, S., Roberts, G.G., Finley, R.L.: DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. Nucleic Acids Research **39**(suppl 1), D736–D743 (2011)
14. Ontology, G.: http://www.geneontology.org/ (Accessed 11 Dec 2013)
15. Pollard, D.: *Drosophila* sequence specific transcription factor binding site matrices, http://www.danielpollard.com/matrices.html (2011). Date accessed: March 2011
16. Przytycka, T.M., Singh, M., Slonim, D.K.: Toward the dynamic interactome: it's about time. Briefings in Bioinformatics **11**(1), 15–29 (2010)
17. Shabalin, A.A., Tjelmeland, H., Fan, C., Perou, C.M., Nobel, A.B.: Merging two gene-expression studies via cross-platform normalization. Bioinformatics **24**(9), 1154–1160 (2008)
18. Sîrbu, A., Ruskin, H.J., Crane, M.: Comparison of evolutionary algorithms in gene regulatory network model inference. BMC Bioinformatics **11**(59) (2010)
19. Sîrbu, A., Ruskin, H.J., Crane, M.: Regulatory network modelling: Correlation for structure and parameter optimisation. In: M. Karim, K. Lee, H. Ling, D. Maroudas, T. Sobh (eds.) Proceedings of The IASTED Technology Conferences (International Conference on Computational Bioscience). Cambridge, Massachusetts (2010)
20. Toledano-Katchalski, H., Nir, R., Volohonsky, G., Volk, T.: Post-transcriptional repression of the drosophila midkine and pleiotrophin homolog miple by how is essential for correct mesoderm spreading. Development **134**(19), 3473–3481 (2007)
21. Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S., Rubin, G.: Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. Genome Biology **3**(12) (2002)
22. Wessels, L.F.A., Reinders, M.J.T., Backer, E.: Robust genetic network modeling by adding noisy data. In: IEEE - EURASIP Workshop on Nonlinear Signal and Image Processing (2001)
23. Xulvi-Brunet, R., Li, H.: Co-expression networks: graph properties and topological comparisons. Bioinformatics **26**(2), 205–214 (2010)