

# Integrating Heterogeneous Gene Expression Data for Gene Regulatory Network Modelling

Alina Sîrbu · Heather J. Ruskin ·  
Martin Crane

Received: date / Accepted: date

**Abstract** Gene regulatory networks (GRNs) are complex biological systems that have a large impact on protein levels, so that discovering network interactions is a major objective of Systems Biology. Quantitative GRN models have been inferred, to date, from time series measurements of gene expression, but at small scale, and with limited application to real data. Time series experiments are typically short, (number of time points of the order of 10), while regulatory networks can be very large, (containing hundreds of genes). This creates an under-determination problem, which negatively influences the results of any inferential algorithm. Presented here is an integrative approach to model inference, which has not been previously discussed, to the authors' knowledge. Multiple heterogeneous expression time series are used to infer the same model, and results are shown to be more robust to noise and parameter perturbation. Additionally, a wavelet analysis shows that these models display limited noise over-fitting within the individual datasets.

**Keywords** gene expression · wavelets · data integration · genetic regulatory networks · complex systems · mathematical modelling

---

This work has been developed under an 'Embark Initiative' grant from the Irish Research Council for Science, Engineering and Technology .

A. Sîrbu

Centre for Scientific Computing and Complex Systems Modelling, School of Computing,  
Dublin City University, Glasnevin, Dublin 9, Ireland  
Tel.: +353-1-7006474  
E-mail: asirbu@computing.dcu.ie

H. J. Ruskin

Centre for Scientific Computing and Complex Systems Modelling, School of Computing,  
Dublin City University, Glasnevin, Dublin 9, Ireland  
E-mail: hruskin@computing.dcu.ie

M. Crane

Centre for Scientific Computing and Complex Systems Modelling, School of Computing,  
Dublin City University, Glasnevin, Dublin 9, Ireland  
E-mail: mcrane@computing.dcu.ie

## 1 Introduction

Modelling gene regulatory networks (GRNs) has received a lot of attention in the literature over the last 10 years [23]. Models range from coarse-grained qualitative models (e.g. Boolean networks) to fine-grained quantitative models (e.g. differential equations) [2]. The latter are typically inferred from time series measurements of gene expression, [30], and provide more detailed information on the dynamics of the complex system, facilitating quantitative simulation and analysis of its behaviour. However, model granularity limits the scale of the analysis, so that, at present, quantitative model inference can be performed only for small scale systems. This is also due to the fact that, in general, since *in vivo/vitro* experiments are costly, the length of time series expression data is limited. This leads to *under-determination* [30], also known as *the curse of dimensionality*. This means that multiple parameter sets are able to reproduce the behaviour seen in the data, and, in consequence, means of discriminating between these are necessary [9].

In order to increase the scale and quality of quantitative GRN modelling, the problem of under-determination has to be tackled. Approaches based on clustering genes before model inference, or reducing connectivity and thus the number of parameters to be inferred can be found in the literature [30]. Further, it is widely recognised that integrative modelling approaches are required to enhance regulatory analysis, [21, 2], and these have started to appear in recent years, mostly for coarse-grained analysis [10, 13] or based on Bayesian models [4]. These integrate expression data with other types of measurements, such as binding affinities or protein interactions, to better discriminate between candidate models. Also, multiple time series datasets from the same single channel platform have been used for linear model inference using singular value decomposition [27]. However, integrating gene expression data from different platforms has been formerly analysed only for *tissue (sample) classification* [6].

Here, an analysis of expression data integration in the context of quantitative *GRN modelling* is presented, using microarray time series datasets from *different platforms*. The hypothesis being investigated is that using such heterogeneous datasets is possible and gives models which are more robust to data and parameter perturbations. Integration helps to capture essential dynamics in the data, without noise over-fitting, (shown by the wavelet analysis described in Section 3.2). While methods can be extended to include other types of expression data, such as Next Generation Sequencing, time series are more difficult to obtain from online databases, due to the novelty of the technique.

## 2 Methods

### 2.1 Model

The analysis presented here has been performed using the S-System model [22]. This is a differential equation system that models the change in gene expression with time using the power-law formalism:

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^n x_j^{g_{ij}} - \beta_i \prod_{j=1}^n x_j^{h_{ij}} \quad (1)$$

Here,  $x_i$  denotes the expression level of gene  $i$ ; the first and second terms represent the synthesis and degradation of mRNA, which are influenced in a positive or negative way by the genes in the network. The rate constants,  $\alpha_i$  and  $\beta_i$ , represent basal synthesis and degradation rate, while  $g_{ij}$  and  $h_{ij}$ , the kinetic orders, indicate the strength of the influence of gene  $j$  on synthesis and degradation of gene  $i$ , respectively. Positive values of  $g_{ij}$  indicate activation of gene  $i$  by gene  $j$ , while negative values indicate repression.

Building an S-System GRN model involves finding the set of parameters  $\{\alpha_i\}$ ,  $\{\beta_i\}$ ,  $\{g_{ij}\}$  and  $\{h_{ij}\}$  that can reproduce the time series used for inference.

### 2.2 Data

Integration analysis has been performed on four distinct datasets representing microarray time series measurements during the Yeast *Saccharomyces Cerevisiae* cell cycle available online: Spellman [26], PramilaS [19], PramilaL [20] and Hasse [18] datasets. Each of these analyse two cell cycles, at different time intervals. The Spellman dataset contains 18 time points sampled every seven minutes, measured using cDNA microarrays, while the PramilaS dataset contains 13 time points, sampled every 10 minutes on Amplicon v1.1 microarrays(cDNA). The PramilaL dataset contains 25 time points sampled every 5 minutes on the same Amplicon platform, and features a dye-swap replicate, which is used in our experiments as a second time series of the same length. The Hasse dataset contains 15 time points, sampled with Affymetrix arrays every 16 minutes, and a replicate that is again used as a separate time series during inference. This results in six time series measurements of the cell cycle sampled at different intervals. The analysis here was performed on a subset of 9 genes known to be involved in the cell cycle, retrieved from KEGG database [3]. These were selected to form a sub-network poorly connected and relatively independent of the rest of the GRN, to facilitate separate analysis.

The data must be normalised for integration to be possible. *Normalisation for noise* has to be performed to remove random and systematic noise introduced at different experimental stages [7]. Additionally, for GRN inference, a *normalisation for scale* needs to be performed. This brings all data in the

same range, to enable the same model parameters to describe the behaviour seen in the different datasets. Typically, normalisation for scale is achieved by standardisation [6]:

$$x' = \frac{x - \bar{x}}{s} \quad (2)$$

Furthermore, when multiple datasets are used for inference, the time spans between consecutive measurements in different datasets have to be measured using the same unit, so a *time span normalisation* is also required. This heavy pre-processing, involving so many stages, is mandatory for the integration to be possible. However, the risk of removing important features during pre-processing exists, as well as introducing false correlations and, in consequence, influencing the resulting GRN model [14].

In this study, the four datasets retrieved from online databases had already been normalised for noise by the authors themselves, but given the different platforms used, the data values had different amplitudes. Consequently, a further scale normalisation has been performed as follows. Firstly, each dataset was standardised and, secondly, the values in all datasets were scaled to the interval  $[0, 1]$  (by subtracting the minimum overall value and then dividing by the maximum value). This was required because of the type of the model used: the S-System model requires positive values for gene expression levels. Additionally, the time spans were modified to bring the cell cycle length to the same level, i.e. 120 minutes. This approach gave good results in our analysis. However, different analyses of normalisation techniques for multi-platform integration [11, 24] have been performed recently, and these may improve results in the context of quantitative GRN modelling. Also, some normalisation techniques, specific to single and dual-channel microarrays, have been extended to be used for both types of microarrays, such a *Loess* or *dChip* [7]. We plan to analyse their effect on data integration in future work.

### 2.3 Inferential Approach

Several inferential algorithms for regulatory network modelling exist in the literature, and we have implemented and used one that has proven performance with real microarray data [25]. This is an evolutionary algorithm based on a hybridisation between Differential Evolution and Hill Climbing local search [16]. Fitness evaluation is based on Akaike’s Information Criterion[1], which uses the MSE between the simulated and real data as a measure of model quality. Model parameters are inferred for each gene at a time, using a *divide-and-conquer* approach. More information on the algorithm can be found in the original paper [16].

In order to analyse performance when moving from one to more datasets, we have split the four time series into two subsets: *inference(training)* and *test* datasets. The *inference* subset has been used during model inference, while the

resulting models have been applied to simulate the *test* series. This bootstrapping approach has been used several times, resulting in thirteen experiments, each using a different combination of datasets for inference. Given the stochastic nature of the evolutionary algorithm, twenty runs have been performed for each of these experiments. All the results presented in this paper are based on the models obtained for gene *CLN2*. This gene was chosen as it is one of the genes that are differentially expressed during different stages of the cell cycle. Similar analysis has been performed with the gene *CLN1*, but results were similar so they are not shown here.

## 2.4 Wavelets

Wavelets [12] are a mathematical tool used for signal processing, which permit a simultaneous time and scale analysis of the signal. At *large scale*, i.e. *low frequencies*, general features of the data can be analysed, while at *small scale*, i.e. *high frequencies*, more detailed aspects are investigated. *Mother* wavelets are functions that satisfy a set of conditions [12], and contracted and dilated versions of these functions, i.e. *wavelets*, are used to decompose the signal into components corresponding to different scales. Continuous wavelet analysis describes the signal as a continuous superposition of wavelets, while discrete wavelet analysis, employed here, as a discrete superposition.

Discrete wavelet decomposition results in a set of coefficients that describe the signal at different scales. The frequency spectra is spanned using smaller and smaller windows, (usually size of windows is decreased twofold from one iteration to another), and coefficients are computed at each level. Level 1 coefficients correspond to the smallest scale, i.e. the highest frequencies in the data, (the upper 50% of the frequency spectrum), level 2 to the next scale (the 25-50% section of the frequency spectrum), and so on. Having  $2^k$  time points in the data,  $2^{k-1}$  level 1 coefficients are computed, for short time windows (total time divided by number of coefficients),  $2^{k-2}$  level 2, for double sized time windows, while the last level,  $k$ , contains 2 coefficients, for large time windows. Each of these coefficients indicates in what amount the current frequency spectrum is present in the signal in the current time window. This results in high time resolution and low frequency resolution at small scale and vice versa at large scale.

In real world applications, noise effects are high frequency, thus visible at small scale in wavelet analysis, while much of the signal is reflected in low frequencies, i.e. large scale. Gene expression time series measurements can be considered as signal levels; we thus performed a wavelet decomposition of the real signal (seen in the four datasets) corresponding to gene *CLN2* in the cell cycle GRN. For this, the signal was resampled using Spline interpolation to generate 32 data points. This was necessary because the number of points required for the wavelet analysis needs to be a power of 2, with  $2^5$  just larger than the number of samples in each time course dataset. In this way we avoided under-sampling for all datasets and loss of information, as well as

over-sampling, to reduce interpolation errors. The Haar[12] wavelet was used for decomposition and resulted in 32 coefficients at 5 scales (levels). The last two scales, 4 and 5, containing 4 coefficients, were combined and labelled as level 4,5 coefficients in this paper. The decomposition was performed using the Matlab toolbox WaveLab [8].

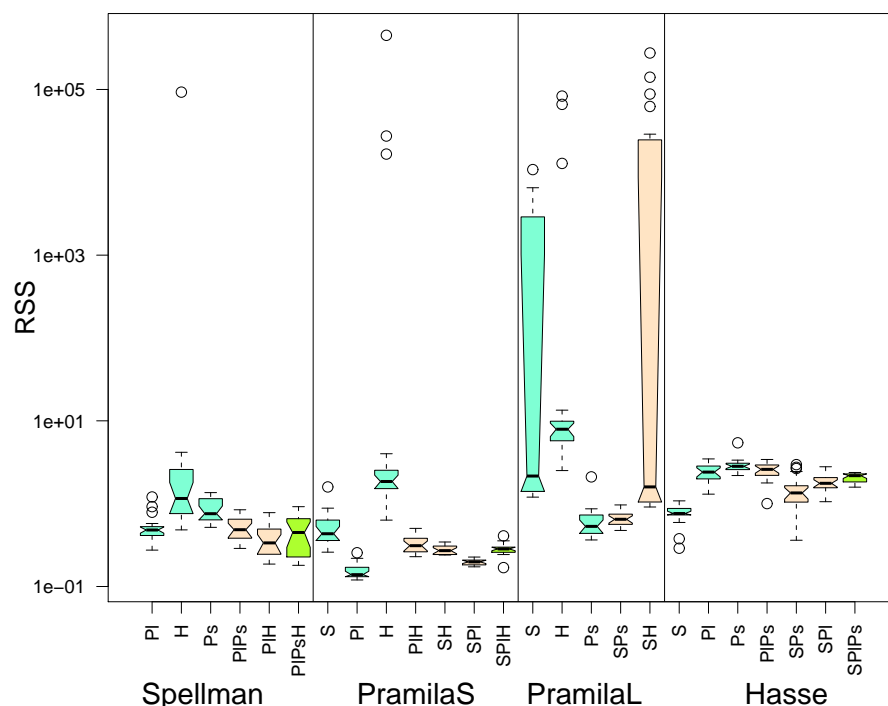
### 3 Results and Discussion

#### 3.1 Performance on Test Datasets

Quantitative integration of cross-platform time-series data is not straightforward, due to platform differences that may interfere with the model which, in consequence, may describe a different process. This is why we have chosen the train/test datasets methodology, to make sure the data is compatible and the models obtained valid.

The results show that for a specific time series used as a test dataset, the data fit of the models obtained from multiple inference datasets is generally better than for those models inferred from one time series, as shown in Figure 1. For instance, in using the Spellman dataset to test models inferred from PramilaL and Hasse together, the figure shows the RSS (Residual Sum of Squares) values to be, on average, lower than when using each dataset separately. This indicates that even though the points added when moving from one to more datasets are from a different source, the resulting model still captures more essential features in the data, (as expected when adding different time points from a single platform), and, in consequence, is better able to describe new test data.

Figure 1 also displays an expected exception: datasets PramilaS and PramilaL describe one another very well, as the performance of the models obtained by using one as inference dataset, and the other as test dataset, is better than when using more datasets. This is, probably, due to the fact that the measurements are performed on the *same platform* and in the *same laboratory* (although at different times), so the datasets lie in the same space and are very close to one another (the measuring time spans also overlap). This explains the behaviour seen for the Spellman and Hasse test datasets: the RSS values for the models inferred from both PramilaL and PramilaS datasets are not significantly different than when using each of these individually. This indicates that, when two datasets are very similar, and the measurement time points overlap, the integration does not significantly improve the RSS criteria. However, we believe that, had the time spans been different, the integration would have brought value to our experiments, as seen for the heterogeneous datasets in this analysis.



**Fig. 1** Performance of models on test datasets. Graphs are displayed as notched boxplots [15], (showing medians and quartile ranges over 20 models for each experiment), using the R module from Wessa.net [28], in order to allow for an analysis of the distribution of RSS values over all models obtained in different runs for each experiment, and to identify statistically significant differences between these. On the  $x$  axis different experiments are represented, using one to three inference datasets as follows: S - Spellman, P1 - PramilaL, Ps - PramilaS, H - Hasse. The four boxes correspond to the four datasets used for test, as labelled.

### 3.2 Wavelet Analysis

A second issue when building GRNs from data is noise handling, as this can influence the resulting model, due to noise over-fitting. Naturally, integrating heterogeneous datasets decreases overall variability, but, in order to avoid over-fitting and model disturbances, the variable elements eliminated have to be those corresponding to noise. This means variability between datasets with respect to main features has to be small enough to facilitate integration. In this section, we will show, using wavelets, that the datasets used in the analysis are compatible, and the models obtained from more datasets display lower noise over-fitting.

Firstly, we compared the coefficients for the four datasets used for inference. For this, the Pearson correlation coefficient [5] was computed at each level, and results are shown in Table 1. Correlation values between coefficients, corresponding to a pair of signals, show how similar the features of the two signals are at different scales. As the values in the table show, all four datasets

**Table 1** Correlation of wavelet coefficients for four gene expression time series (Spellman - S, PramilaS - Ps, PramilaL - Pl, Hasse - H). Level 1 corresponds to the smallest scale, i.e. highest frequencies, while level 4,5 to lowest frequencies.

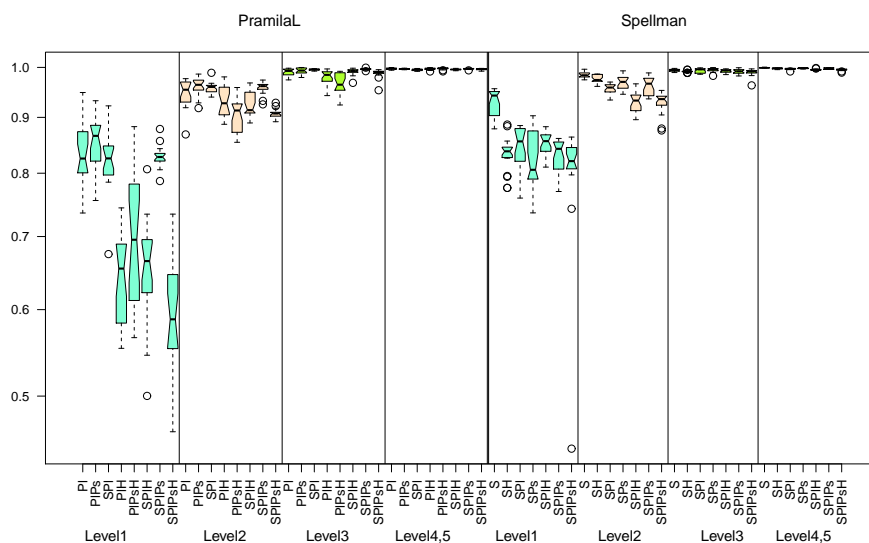
	Level 1			Level 2			Level 3			Level 4,5		
	Ps	Pl	H	Ps	Pl	H	Ps	Pl	H	Ps	Pl	H
S	.582	.742	.087	.950	.890	.143	.907	.996	.844	.994	.998	.966
Ps		.283	.078		.902	.167		.939	.707		.991	.938
Pl			.156			.316			.828			.975

are very similar at levels 3 and 4,5 (corresponding to low frequencies, i.e. real features), while at levels 1 and 2, (i.e. noise effects), most correlations are very low, indicating heterogeneity of noise in the datasets. The high correlation in essential features indicates that the datasets are compatible for integration. The low correlations of the Hasse dataset with all other datasets at the first two levels show that this dataset is very different from the others, and this may be because of the difference in microarray platform (one channel vs. two channels). Also, correlations indicate many similarities, even at levels 1 and 2, between the three datasets measured on two channel microarrays, indicating these lay in the same space. However, this was visible in the experiments presented in Section 3.1 only for datasets PramilaL and Pramilas, confirming our hypothesis that, even when datasets describe very similar features in terms of dynamics, non-overlap of measured time points means the integration improves RSS on test datasets, (as expected, when adding more time points to the analysis).

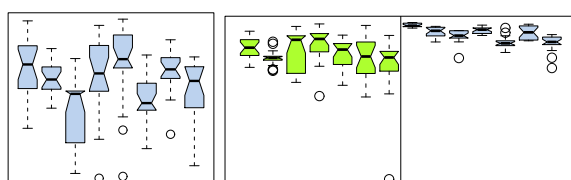
To analyse noise over-fitting, wavelet coefficients for the simulated signals by the models inferred in different runs were also computed. The coefficients obtained at each level for each model were compared to those in the initial data, by computing the same Pearson correlation coefficient. The values obtained are displayed in Figure 2 for the Spellman and PramilaL datasets.

Results show that correlation values at levels 1 and 2, (corresponding to high frequencies, i.e. noise), decrease when adding more time series to the inference process, while those corresponding to real features in the data, (levels 3 and 4,5), are stable. This proves that using heterogeneous time series reduces noise over-fitting, (as different types of noise exist in the data, and it is very difficult for the inferential algorithm to over-fit all types), while main features in the data are maintained. For instance, for the Spellman dataset, when the model was inferred only from the Spellman time series, the correlations are high at all levels (including those corresponding to noise), indicating over-fitting of noise. However, when adding other datasets to the inference and comparing again with the initial Spellman data, correlations decreased for high frequency levels, while for low frequencies they remained high.





(a) PramilaL and Spellman all levels.



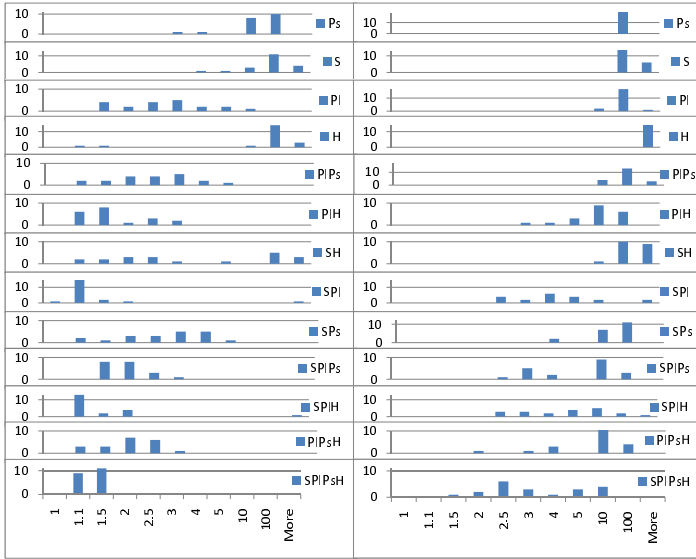
(b) PramilaL level 1 enlarged. (c) Spellman levels 3 and 4,5 enlarged.

**Fig. 2** Correlation of wavelet coefficients for datasets PramilaL and Spellman at different levels. Image 2(a) shows all levels for both datasets, while 2(b) and 2(c) show enlarged images of three levels for which details are poorly visible in the main image, due to scale differences. When moving from one to more inference datasets, correlations at levels 1 and 2 (high frequencies) decrease, while those at levels 3 and 4 remain very high, indicating less noise over-fitting, while maintaining important features in the data.

### 3.3 Robustness Analysis

The transcriptional process is stochastic in nature [23], and real GRNs are robust to relatively small changes in expression values. Also, it has been argued that quantitative models should also be robust to small parameter perturbations [9], as the connectivity of the network is the main factor influencing causal effects [2]. By integrating heterogeneous data, with different types of noise present, more robust models should be obtained. This has been tested in this paper by performing a sensitivity analysis for noise and parameter perturbation.

Firstly, models obtained in different runs were analysed on data containing Gaussian noise, which was added to the initial datasets used for inference, similar to [17]. Although, in reality, the distribution of the noise may be different,

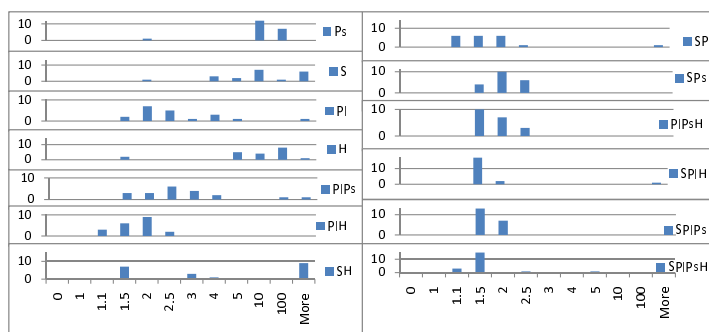


**Fig. 3** Sensitivity to noise. Histograms show RSS ratios over 20 models for different experiments, ranging from one (top) to four (bottom) inferential datasets, under two noise levels (standard deviations 0.01 - left - and 0.05 - right). A clustering to the left of the ratios can be seen when moving from top to bottom under both noise levels.

this analysis gives a good indication on the robustness of the models to given percentage level of perturbations in the data. The resulting simulation time series were compared to the initial data and RSS values compared to those obtained with data without added noise (a ratio value was computed). Figure 3 plots histograms of RSS ratios for the 20 models in each experimental run, under two noise levels (deviation of 0.01 and 0.05). When moving from experiments with one to more inference datasets, more models maintain a good RSS value (ratios around 1), indicating the models obtained using integrated data are more robust to noise.

Secondly, a parameter sensitivity analysis was performed. For this, for each model inferred, individual parameters were slightly modified and the time series simulated and compared. A RSS ratio against initial RSS (obtained by the unmodified model) was computed, analogous to the noise sensitivity analysis. Figure 4 shows histograms of ratios obtained for the 20 models in each experiment, with perturbations of  $\pm 1\%$  of initial search interval, (0.2 in our case). Again, a clustering of the RSS ratios to the left, (lower values), can be seen when moving from experiments using one dataset to those using three or four datasets, indicating better resilience to parameter perturbations for the latter.

It is important to note that even similar datasets, such as PramilaL and PramilaS, add robustness when used together. For instance, Figures 3 and 4 show that models inferred from all four datasets are more robust to both types of perturbations than those inferred from three datasets by eliminating



**Fig. 4** Sensitivity to parameter perturbations. Histograms of RSS ratios are plotted for 20 models obtained in each experiment. Ratios are smaller for models obtained from multiple datasets (bottom right), compared to those from one dataset only (top left).

PramilaS. This is due to the stochasticity of gene expression: even if the series describe, in essence, the same process, small variability exists and forces the resulting model to handle this. This is similar to adding noisy replicates to the inferential process, which has been previously proven to give more robust models [29].

#### 4 Conclusions and Future Work

We have shown that integration of multiple time series for GRN quantitative model inference is possible and can result in improved models. We have inferred GRN S-System models from four gene expression datasets measured on different platforms, and analysed these and their simulated data. A robustness analysis proved models obtained from multiple datasets were more resilient to both noise in the data and parameter perturbations. Additionally, a wavelet decomposition of signals corresponding to gene CLN2 has been performed, and results proved that over-fitting of noise was reduced for models inferred from heterogeneous time series.

In this analysis, normalisation for noise was performed by the authors, and their normalised data were used. However, this step is very important in microarray data pre-processing, as it can influence resulting models. A further analysis on how different normalisation techniques influence the integration process is required, and will be performed in future work. Also, an analysis of the entire GRN, as opposed to single gene models, is required to further validate results. Additionally, tests on other types of gene expression data are planned, (e.g. RNA-seq data), to demonstrate that results obtained can be extended to other cases.

## 5 List of Abbreviations

GRN - Gene Regulatory Network, mRNA - messenger RNA, cDNA - complementary DNA, RSS - Residual Sum of Squares, S - Spellman dataset, Pl - PramilaL dataset, Ps - PramilaS dataset, H - Hasse dataset.

## References

1. Akaike, H.: A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19**(6), 716–723 (1974)
2. Alvarez-Buylla, E.R., Benitez, M., Davila, E.B., Chaos, A., Espinosa-Soto, C., Padilla-Longoria, P.: Gene regulatory network models for plant development. *Current Opinion in Plant Biology* **10**(1), 83 – 91 (2007). DOI DOI: 10.1016/j.pbi.2006.11.008. URL <http://www.sciencedirect.com/science/article/B6VS4-4MGVPM8-1/2/7001658f1157dce0838fa1c333a96043>
3. Aoki-Kinoshita, K.F., Kanehisa, M.: Gene annotation and pathway mapping in KEGG. *Methods in molecular biology (Clifton, N.J.)* **396**, 71–91 (2007). URL <http://view.ncbi.nlm.nih.gov/pubmed/18025687>
4. Bernard, A., Hartemink, A.J.: Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. In: *Proceedings of Pacific Symposium on Biocomputing*, pp. 459–470. World Scientific, Duke University, Dept of Computer Science, Durham, NC 27708, USA. (2005). URL <http://view.ncbi.nlm.nih.gov/pubmed/15759651>
5. Boslaugh, S., Watters, P.A. (eds.): *Statistics in a Nutshell*. O’Reilly Media, Inc. (2008)
6. Cheng, C., Shen, K., Song, C., Luo, J., Tseng, G.C.: Ratio adjustment and calibration scheme for gene-wise normalization to enhance microarray inter-study prediction. *Bioinformatics* **25**(13), 1655–1661 (2009). DOI 10.1093/bioinformatics/btp292. URL <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/13/1655>
7. Do, J.H., Choi, D.K.: Normalization of Microarray Data: Single-labeled and Dual-labeled Arrays. *Molecules and Cells* **22**(3), 254–261 (2006)
8. Donoho, D., Maleki, A., Shahram, M.: Wavelab and reproducible research. [http://www-stat.stanford.edu/~wavelab/Wavelab.850/index\\_wavelab850.html](http://www-stat.stanford.edu/~wavelab/Wavelab.850/index_wavelab850.html) (1995). URL [http://www-stat.stanford.edu/wavelab/Wavelab.850/index\\_wavelab850.html](http://www-stat.stanford.edu/wavelab/Wavelab.850/index_wavelab850.html)
9. Fomekong-Nanfack, Y., Postma, M., Kaandorp, J.: Inferring drosophila gap gene regulatory network: a parameter sensitivity and perturbation analysis. *BMC Systems Biology* **3**(1), 94 (2009). DOI 10.1186/1752-0509-3-94. URL <http://www.biomedcentral.com/1752-0509/3/94>
10. Huttenhower, C., Mutungu, K.T., Indik, N., Yang, W., Schroeder, M., Forman, J.J., Troyanskaya, O.G., Collier, H.A.: Detailing regulatory networks through large scale data integration. *Bioinformatics* **25**(24), 3267–3274 (2009). DOI 10.1093/bioinformatics/btp588. URL <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/24/3267>
11. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat* **8**(1), 118–127 (2007). DOI 10.1093/biostatistics/kxj037. URL <http://biostatistics.oxfordjournals.org/cgi/content/abstract/8/1/118>
12. Kaiser, G.: *A Friendly Guide to Wavelets*. Birkhauser (1994). URL <http://dx.doi.org/>
13. Kundaje, A., Xin, X., Lan, C., Lianoglou, S., Zhou, M., Zhang, L., Leslie, C.: A predictive model of the oxygen and heme regulatory network in yeast. *PLoS Comput Biol* **4**(11), e1000224 (2008). DOI 10.1371/journal.pcbi.1000224
14. Lim, W.K., Wang, K., Lefebvre, C., Califano, A.: Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics* **23**(13), i282–288 (2007). DOI 10.1093/bioinformatics/btm201. URL <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/13/i282>
15. McGill, R., Tukey, J.W., Larsen, W.A.: Variations of box plots. *The American Statistician* **32**(1), 12–16 (1978)

16. Noman, N., Iba, H.: Inference of genetic networks using S-system: information criteria for model selection. In: GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation, pp. 263–270. ACM, New York, NY, USA (2006). DOI <http://doi.acm.org/10.1145/1143997.1144043>
17. Noman, N., Iba, H.: Inferring gene regulatory networks using differential evolution with local search heuristics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **4**(4), 634–647 (2007)
18. Orlando, D.A., Lin, C.Y., Bernard, A., Wang, J.Y., Socolar, J.E., Iversen, E.S., Hartemink, A.J., Haase, S.B.: Global control of cell-cycle transcription by coupled cdk and network oscillators. *Nature* **453**(7197), 944–947 (2008). DOI 10.1038/nature06955. URL <http://dx.doi.org/10.1038/nature06955>
19. Pramila, T., Miles, S., GuhaThakurta, D., Jemiolo, D., Breeden, L.L.: Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. *Genes and Development* **16**(23), 3034–3045 (2002). DOI 10.1101/gad.1034302. URL <http://genesdev.cshlp.org/content/16/23/3034.abstract>
20. Pramila, T., Wu, W., Miles, S., Noble, W.S., Breeden, L.L.: The Forkhead transcription factor HCM1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes and Development* **20**(16), 2266–2278 (2006). DOI 10.1101/gad.1450606. URL <http://genesdev.cshlp.org/content/20/16/2266.abstract>
21. Przytycka, T.M., Singh, M., Slonim, D.K.: Toward the dynamic interactome: it's about time. *Briefings in Bioinformatics* **11**(1), 15–29 (2010). DOI 10.1093/bib/bbp057. URL <http://bib.oxfordjournals.org/cgi/content/abstract/11/1/15>
22. Savageau, M.A.: Introduction to S-Systems and the underlying power-law formalism. *Mathematical and Computer Modelling* **11**, 546–551 (1988). URL <http://hdl.handle.net/2027.42/27576>
23. Schlitt, T., Brazma, A.: Current approaches to gene regulatory network modelling. *BMC Bioinformatics* **8**(Suppl 6), S9 (2007). DOI 10.1186/1471-2105-8-S6-S9. URL <http://www.biomedcentral.com/1471-2105/8/S6/S9>
24. 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). DOI 10.1093/bioinformatics/btn083. URL <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/9/1154>
25. Sirbu, A., Ruskin, H.J., Crane, M.: Comparison of evolutionary algorithms in gene regulatory network model inference. *BMC Bioinformatics* **11**(59) (2010)
26. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9**(12), 3273–3297 (1998). URL <http://www.molbiolcell.org/cgi/content/abstract/9/12/3273>
27. Wang, Y., Joshi, T., Zhang, X.S., Xu, D., Chen, L.: Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics* **22**(19), 2413–2420 (2006). DOI 10.1093/bioinformatics/btl396. URL <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/19/2413>
28. Wessa, P.: Notched boxplots (v1.0.5) in free statistics software (v1.1.23-r4) (2008). URL [http://www.wessa.net/rwasp\\_notchedbox1.wasp/](http://www.wessa.net/rwasp_notchedbox1.wasp/)
29. 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)
30. Xu, R., Hu, X., Wunsch D.C., I.: Inference of genetic regulatory networks from time series gene expression data. *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on* **2**, 1215–1220 (2004)