Noname manuscript No.
(will be inserted by the editor)

Reconstructing Biological Gene Regulatory Networks: Where Optimization Meets Big Data

Spencer Angus Thomas · Yaochu Jin

Received: date / Accepted: date

Abstract The importance of 'big data' in biology is increasing as vast quantities of data are being produced from high-throughput experiments. Techniques such as DNA microarrays are providing a genome-wide picture of gene expression levels, allowing us to investigate the structure and interactions of gene networks in biological systems. Inference of gene regulatory network (GRN) is an underdetermined problem suited to Metaheuristic algorithms which can operate on limited information. Thus GRN inference offers a platform for investigations into data intensive sciences and large scale optimization problems. Here we examine the link between data intensive research and optimization problems for the reverse engineering of GRNs. Briefly, we detail the benefit of the data deluge and the study of ALife for modelling GRNs as well as their reconstruction. We discuss how metaheuristics can solve big data problems and the inference of GRNs offer real world problems for both areas of research. We overview some current reconstruction algorithms and investigate some modelling and computational limits of the inference processes and suggest some areas for development. Furthermore we identify links and synergies between optimization and big data, e.g., dynamic, uncertain and large scale optimization problems, and discuss the potential benefit of multi- and many-objective optimization. We stress the importance of data integration techniques in order to maximize the data available, particularly for the case of inferring GRNs from microarray data. Such multi-disciplinary research is vital as biology is rapidly becoming a quantitative, data intensive science.

Spencer Angus Thomas
Department of Computing
University of Surrey
Guildford, Surrey, GU2 7XH, United Kingdom
Tel.: +44 1483 68 6056
E-mail: s.thomas@surrey.ac.uk

Yaochu Jin
Department of Computing
University of Surrey
Guildford, Surrey, GU2 7XH, United Kingdom
Tel.: +44 1483 68 6037
E-mail: yaochu.jin@surrey.ac.uk

Keywords Metaheuristics · Evolutionary Algorithms · Data-Driven Optimization · Gene Regulatory Network Reconstruction · Big Data · Data Science

1 Introduction

Biological systems can be modelled using gene regulatory networks (GRNs), where a group of genes influence each other's dynamic behaviour. GRNs are the most important organization level within a cell [30], and are the focus of much research in the growing field of systems biology [49]. The building blocks of gene networks are not well known [52, 116, 129], however the role of each gene can be better understood by investigating their interactions and topology within GRNs [46]. Systems biology can broaden our knowledge about networks that are responsible for basic biological functions and robustness, and the causes of their breakdowns leading to disease states [76]. How cellular systems are formed from the interactions between genes, proteins and small molecules is a major challenge for biology [26].

1.1 Modelling

There are several methods for modelling GRNs. Here we briefly cover Boolean, Bayesian and differential equation models, for details of common techniques the reader is referred to [12, 46, 54, 70, 134]. Logic models, the most fundamental of which are the Boolean networks [73, 102], are a popular choice as they can give information about the network topology and are relatively simple to analyse [102]. For a Boolean model of a two gene system, each gene can be either active (1) or inactive (0) [137] and interactions can be modelled using *IF* statements. For the GRN given in Fig. 1, a simple Boolean network would model the regulations as

$$g_1 = \begin{cases} 0 & \text{if } g_2 \text{ is } 1 \\ 1 & \text{if } g_2 \text{ is } 0 \end{cases} \quad \text{Repressor,}$$
$$g_2 = \begin{cases} 0 & \text{if } g_1 \text{ is } 0 \\ 1 & \text{if } g_1 \text{ is } 1 \end{cases} \quad \text{Activator,} \qquad (1)$$

where g_1 activates itself and g_2, while g_2 represses g_1. This leads to a flipping of each of the genes from the inactive (0) state to the active (1) state due to the repression of g_1 by g_2 and demonstrates the importance of Boolean networks in understanding steady states and robustness in GRN [73].

It is possible to model a GRN using Bayes' theorem and this technique is commonly used to reconstruct biological networks [28, 46, 123, 149, 161]. Bayesian networks are directed acyclic graphs (DAG) that also contain the probabilistic dependencies between two variables [17], in this case genes. It is possible to add directionality to the network through scoring metrics such as the posterior probability [27] as directionality in biological networks is important. Bayesian networks can be static or dynamic [14], where dynamic networks can provide better data recognition than static networks but increase computational run time [40].

More detailed models include ordinary differential equation (ODE) models, which are able to model the dynamics behaviour of each gene in the network, are commonly used [25, 120, 121, 143, 160, 162]. This increase in knowledge of the system, i.e. how the expression of each gene varies over time, comes at an increased cost in model complexity and computational run time as it requires an ODE solver. In general for an N gene network the dynamic behaviour can be modelled as $dx_i/dt = f_i(x_1, x_2, \ldots, x_N)$, where $f_i, i = 1, 2, \cdots, N$ represents the regulation between network genes, and is commonly modelled as a Hill function [71, 144–146, 153]. Hill functions are non-linear equations that are derived from Michaelis-Menten enzymatic kinetics [4, App. A]. A possible ODE model of the GRN in Fig. 1 using Hill functions and summation logic to combine the auto regulation of g_1 with the repression from g_2 would take the form

$$\begin{aligned} \dot{g_1} &= \tfrac{w_1}{2}\left(\tfrac{\beta_1}{1+(\phi_1/g_2)^n} + \tfrac{\beta_1 g_1^n}{\phi_1^n + g_1^n}\right) - \gamma_1 g_1 \\ \dot{g_2} &= w_2 \tfrac{\beta_2 g_1^n}{\phi_2^n + g_2^n} - \gamma_2 g_2 \ , \end{aligned} \tag{2}$$

where w_i is the interaction weight, β_i is the maximum activation, ϕ_i is the threshold for the interaction, γ_i is the degradation of the protein from gene $i, i = 1, 2$, and n is the Hill coefficient. Non-linear models are favoured as interactions in nature, such as gene regulation, often have non-linear characteristics in their behaviour [94, 150].

Another common ODE modelling technique is the S-System [49, 61, 106, 107, 110, 122]. This is a power law formalism, which in general is

$$\dot{x_i} = \alpha_i \prod_{j=1}^{N} x_j^{g_{i,j}} - \beta_i \prod_{j=1}^{N} x_j^{h_{i,j}} \ . \tag{3}$$

For the GRN in Fig. 1 the S-System model would be

$$\begin{aligned} \dot{x_1} &= \alpha_1 x_1^{g_{1,1}} - \beta_1 x_1^{h_{1,1}} x_2^{h_{1,2}} \\ \dot{x_2} &= \alpha_2 x_1^{g_{2,1}} - \beta_2 x_2^{h_{2,2}} \ , \end{aligned} \tag{4}$$

where α_i and β_i are the production and degradation weights. Here we can see that x_1 self-activates in the first term, while the second term combines the repression by g_2 as well as the degradation of g_1's mRNA. For x_2 we have a simpler equation where the first term is for the activation from g_1 and the second term is for the degradation of the mRNA from g_2.

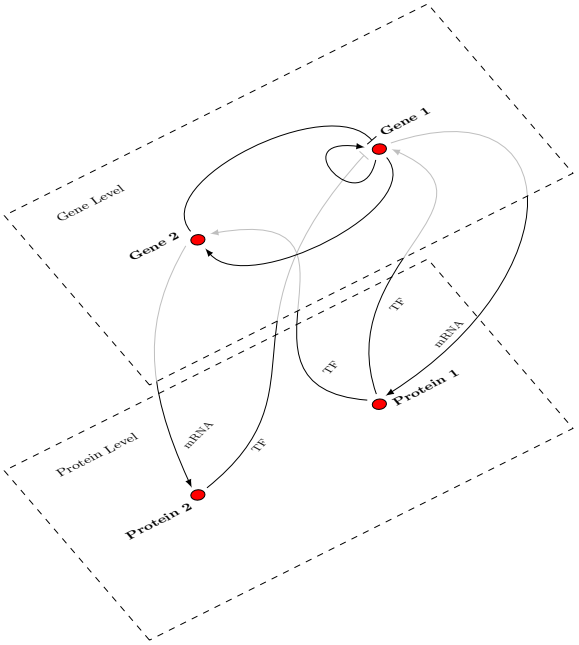

Fig. 1: A gene regulatory network model showing how regulation on the gene level corresponds to interactions with the protein level via mRNA and Transcription Factors (TF)

Other forms of models, such as informatics models, are based on Pearson's correlation [37] or mutual information [29, 99, 111, 163] and use expression profiles to determine the likelihood of a connection between two genes. Other models based on the state of a system include Petri Nets [73] and state charts [39].

1.2 Reconstruction

The reconstruction of biological GRNs is one of the most complex tasks in bioinformatics [102] as GRNs are not fully understood [141]. It is possible to use expression data to reconstruct gene regulatory networks [90], which is an active area of research [49] and one of the most important challenges in systems biology [91]. The reverse engineering of GRNs therefore can serve as an intermediate step between systems biology and bioinformatics [46]. Current high-throughput experiments can provide genome-wide gene expression measurements and an active area of research is the inference of gene networks from this data [49]. Despite experimental advances in data collection techniques, significant costs lead to limited availability for fine grain time series data for a given network. Penfold and Wild [113] noted that for microarray time series data for 3 replicates each with 25 time points costs in the region of £30,000 (over $45,000). Furthermore, specific growth conditions for many organisms mean that much of the data is heterogeneous and can not necessarily be used together. This has led to the underdeterminism of such problems, often referred to as 'the

curse of dimensionally', where there is insufficient time series data available to statistically reconstruct large networks [49, 73, 134].

1.3 Optimization

As biological networks are often large [49], particularly for more complex organisms, sophisticated reconstruction techniques are required. One such technique is the use of optimization algorithms to reconstruct the biological networks based on data which are often noisy and incomplete [7, 49], which is a general problem in biology [153]. Metaheuristic optimization algorithms have the advantage of not requiring detailed prior knowledge of the system, but only an evaluation of potential solutions, and are a powerful tool for modelling complex problems in biology [152]. In order to fully reconstruct a GRN, one must identify both the topology and parameterization of the connections, resulting in a vast search space for the problems and thus making optimization algorithms an attractive method.

1.4 Big Data

Big data is characterised by increasing volume, variety and velocity of data generated [86], see Fig. 2, rather than simply running repeats of the same experiments producing replicate data sets. Recently data veracity (uncertainty and reliability) [136] is also used with the others to form what is known as the 'Four V's of big data'. According to IBM, "Every day, we create 2.5 quintillion bytes of data - so much that 90% of the data in the world today has been created in the last two years alone" [62]. This is 2.5 x 10^{18} bytes per day and results in somewhere in the region of 10^{21} bytes of data in the world today, which is comparable with the mass of the Moon (73.5 x 10^{21} kg) [154].

With this in mind, reconstruction of GRNs poses a biological big data problem and provides a platform for biologists and computer scientists to address this problem from the optimization point of view. The volume of data is ever increasing with developments in next generation sequence techniques such as high-throughput experiments. Analytical and computational developments will eventually allow us to collect, process and analyse data in real time increasing the data velocity. Improvements in next generation sequencing techniques will reduce noise and errors in measurements and help address the veracity of the data. With so many potential experimental conditions and measurement combinations, it is clear that the variety of data will grow with time. However, increasing data variety means that GRN reconstruction techniques must utilize data integration methods in order to build realistic biological models. The need for data integration and 'curation' [16] will increase rapidly as the data deluge pushes biology towards the 'Fourth Paradigm' [55] as data intensive science.

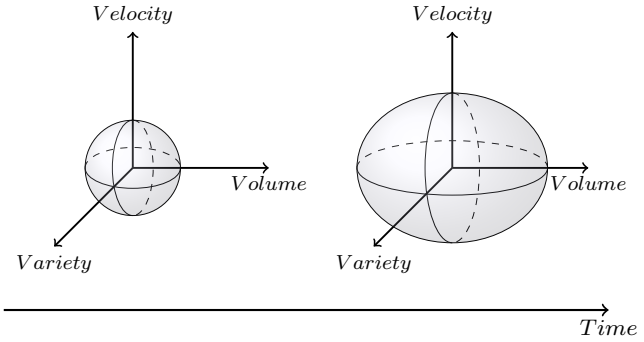

Fig. 2: How data is changing: high-throughput experiments and improvements in data storage will lead to significant increases in the volume and velocity of data over the next few years. Data variety will also increase with additional types of experimental data being stored, however this will likely be at a lower rate the others. Veracity is another aspect of big data though not shown here.

1.5 Optimization and Big Data

Gene expression data is growing at an increasing rate and the fields of biology, systems biology and bioinformatics are entering the Fourth paradigm. As a consequence these fields will have to start dealing with the big data problem and develop universal data collection, curation, storage and retrieval techniques in order to develop our understanding of GRNs during the data deluge. As mentioned, the reconstruction of GRNs from gene expression data is a platform for studying optimization techniques and principles for real world problems. Due to the large search space, sparse and incomplete data, and potentially complex fitness landscapes, these reverse engineering problems are well suited to metaheuristic techniques. Therefore using gene expression data to reconstruct GRNs is both an optimization problem and a big data problem, and can be thought of as a 'Big Optimization Problem'.

1.6 GRNs and ALife

Artificial life (ALife) consists of three domains including; 'soft' ALife, where life-like behaviour is simulated, 'hard' ALife, where life-like systems are implemented, and 'wet' ALife, where biochemical substances are used to synthesize living systems [15]. GRNs are mainly in the 'soft' ALife domain as they can used *in silico* to model each of the levels of transcriptional regulation as described by Babu et al. [8]. They can also be used to simulate phenomenon observed in biology, such as duplication [84], evolvability [**?**], connectivity preference [145], robustness [140], fragility [85], motifs and modularity [129]. Computational modelling can provide insight into biological systems with *in silico* experiments providing predictions that can be tested *in vitro* and *in vivo* [78].

The study of GRNs can be divided into two areas, synthesis and analysis, and reconstruction. The former links GRNs with ALife and researchers in this field are improving our fundamental understanding of biological

properties. The latter is the focus of this paper and is the platform in which we will link optimization and big data. Although ALife and big data may not be linked directly, they are both associated with GRNs and progression on the elementary level, i.e. ALife, will improve understanding and modelling at higher levels, i.e. reconstruction from data.

The rest of the paper is structured as follows; Section 2 provides a brief overview of various modelling methods for biological network reconstructions. In Section 3 we outline some metaheuristic methods for solving optimization problems such as biological networks. We discuss the role of Big Data in biological networks in Section 4. The role of GRNs in artificial life is briefly discussed in Section 5 along with the importance of ALife, alongside big data, in GRN inference. Next we given an overview of some existing reconstruction algorithms in Section 6. Some areas for computational development, particularly in the area of parallelization, are discussed in Section 7. Finally we provide a summary and conclusions on the topic in Section 8.

2 Modelling Networks

2.1 Models

Although GRNs are an abstract modelling method that can be applied to a number of problems from neural networks to ecological food webs, they can also be used to model transcriptional networks [103]. In their abstract form GRNs are nodes that interact directly or indirectly, and the form of this interaction is also abstract and can be problem specific. There are various methods one can use in order to model GRNs such as, static or dynamic, continuous or discrete, linear or non-linear, deterministic or stochastic models [46]. It is possible to model various levels of biological activity from gene regulation and protein interaction to metabolic and biochemical reaction [36]. Figure 1 shows a simplified two gene system which models the gene and protein interaction levels only. This indicates that for a large number of genes even a simplified model can quickly become complicated.

2.2 Dealing With Data

Each data set can be either experimentally measured or synthetically generated from a target network, which is usually the goal of the reconstruction. Synthetic, or artificial data sets are used because the limited availability and cost of experimental data. However it also allows comparisons between the predicted network and the actual known network to be made and thus assess the reliability and performance of the reconstruction method.

Due to the complexity of experimental data sets, artificial data is not a realistic representation of biological data [46], thus varying levels of noise are often added to the data to make them more 'realistic' [46, 132], as well as to further test the algorithms performance [110]. When using artificial data, it is also possible to produce many time points and replicate data sets to aide the reconstruction process [12, 46, 113, 132], which are seldom available for real networks.

Those dealing with real data sets are often required to maximise the amount of data through interpolation of the available data [131]. This has the added benefit of giving constant time steps for the data points, as experimental time series may have varying time intervals in between the measurements [74].

For a large network it is possible to cluster genes with similar expression profiles together to reduce the size of the network and reduce the dimensionality of the problem. This method has been used by several researchers [36, 58, 147], however this adds the need for a clustering process and a method of combining expression profiles before the optimization stage creating additional overhead and potential errors. Although this technique does fit with some biological observations, e.g. sparsity [90] and small world networks, it is still a simplification.

Additional information can be taken from other types of data sets, such as from knockout and perturbation experiments. The former removes a gene from a genome, known as a null or mutant strain, and compares expression levels of the genes with a 'wild type' organism [3]. This process is known as differential gene expression and can give significant insight into which genes are in the same network by observing genes affected by the deletion. The latter gives finer detailed information into the interconnectivity in the network by varying the state of a particular gene to observe any changes in other genes in the network [12].

2.3 Topological Networks

Topological models are of particular importance to our understanding of the behaviour of GRNs due to the modularity of biological systems and the functions of these modules such as AND and OR gates for time delays and robustness [5, 93, 103, 135]. Topological models can also help identify auto-regulation of genes, which are known for their functional roles such as decreasing response time or enhanced variation in expression levels [5].

Swain et al. [141] illustrated the importance of topology by using a caterpillar and butterfly analogy, where the two insects contain the same genes, the connectivity of which is changed during the crystalline phase resulting in the physical difference between the two. In some cases the topology of the network is more important than the parameterization as the structure can determine the dynamic behaviour of the network [49, 141]. The importance of topology over parameterization is also present in more complex models, such as the Drosophila segment polarity network. This model contains 48 free parameters, which when randomised, each had a 90% chance of being compatible with the desired behaviour regardless of parameter magnitude or range [31]. The authors observed the desired dynamics approximately 1 in every 200 runs, much more frequent than at random.

The main issue with topological models is the lack of a quantitative metric for comparison between models

[153]. One can use measures such as specificity, sensitivity, precision and recall, however for the practical case of an unknown topology, theses measures are useless. However for competing models of the same unknown system the fit to the experimental data can be used as a measure of model quality. For models of similar fits to the data, the simpler model, i.e. fewer nodes and/or connections, is preferred as it is easier to understand and less prone to over fitting [73], i.e. Occam's Razor.

2.4 Parameterized Networks

The parameterization of a network is also important as it allows us to investigate the modelling of the connections within a GRN. The difference between a simple linear connection and a more complex non-linear connection between genes could significantly affect their dynamical behaviour. In [63] Ingram and coworkers found that even for a relatively simple motif, a bi-fan, network dynamics vary greatly for different connection types and parameter sets. The authors demonstrated that this simple topology can be tuned to give a desired output and therefore a general statement about a network motif's structure is not sufficient to determine functionality. Gonze [45] observed regions in the parameter space that determined the dynamic behaviour of a fixed network structure and determined the bifurcation values for changing the network dynamics. Similarly in [144] the authors observed different dynamic behaviour for a fixed network by varying the parameters of one of the connections.

2.5 Combining Topology and Parameterization

In [145] it was observed that for a given topology the same behaviour was observed for multiple parameter values, however by reversing the direction of a single connection the network appears to be dependent on parameterization. Here the authors also observed a 'weak' bifurcation point in a single parameter that caused a change in the dynamic behaviour of the network in most cases. Thomas and Jin [146] found that a given topology was capable of producing oscillatory network dynamics, however only for certain parameter sets, and produced trivial dynamics the majority of the time. This, along with somewhat conflicting conclusions from Section 2.3 and Section 2.4 indicates that the dynamic behaviour of a network is influenced by both topology and parameterization.

3 Metaheuristic Methods

Metaheuristic methods can be used to solve difficult optimization problems with little or no prior knowledge. These are able to solve underdetermined problems [104, 134] such as the reverse engineering of GRNs. Although metaheuristic searches do not always yield the most optimal solution, they can provide an acceptable solution given the problem constraints. As many metaheuristics are stochastic, it is possible to average results over numerous simulations or to find the optimum solution. The general outline of a metaheuristic algorithm is given in Fig. 3.

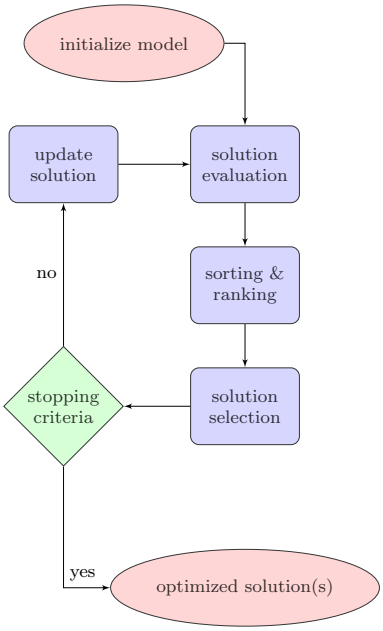

Fig. 3: General process of a metaheuristic algorithm.

Reverse engineering GRNs is a difficult task hindered by the complexity of biological networks [42]. Both the reconstruction of the network topology and the interactions between network nodes are suited to optimization as a complex real world problem with a large search space containing many local optima [7]. Leclerc [90] stated that if biological networks are optimal, the search space for robust functional networks may contain highly adaptive peaks separated by great distances corresponding to different network topologies. For such an optimization problem, the ability to escape local optima is vital to ensure convergence to a biologically plausible optimal solution. Several optimization techniques have been used for reconstructing biological networks, however they are limited by the amount of data available and the high dimensionality of the problem [118], as well as computational power required for large networks.

3.1 Nature Inspired Optimization Algorithms

Evolutionary algorithms (EAs) are able to deal with large search spaces [102] and complex fitness landscapes, such as Fig. 4, and are therefore well suited to network reconstruction problems [141]. Several EAs have been used to infer GRNs from differential evolution [109, 110], a genetic algorithms [74], an evolution strategy [137], genetic programming and particle swarm optimization [23].

3.2 Elitism

The role of elitist selection in evolutionary optimization is a debated issue due to its ability to aide algorithm convergence [69, 144–146] but also leads to local, rather than global, optimal solutions [19]. Reverse engineering biological networks in general have a large optimization search

space as a result of the systems complexity, and even the smallest known genome can contain 200,000 interactions, after making many false simplifications [49]. If you modelled a cell through all its significant interacting constituents, the resulting complexity would be 'daunting' [75]. Due to such large search spaces, many researchers have used elitism in their reconstruction algorithms to aide the evolutionary search and reduce the computation time [61, 92, 102].

3.3 The Number of Objectives

3.3.1 Single Objective

Optimization algorithms are designed to solve a problem based on an objective function. If there is only one objective, such as error minimization then this is a single objective problem. A common objective for biological network reconstruction is simply to minimise the error between the model output and the data [110, 120, 134, 149, 157], which can be applied to both real and Boolean networks [49]. This method can work well for small networks, however, this can lead to over fitting and many false positive connections if there is no constraint on number of connections between nodes. This becomes a problem of larger networks, not only for biological relevance but it also increases computation time dramatically. Other objective functions have been suggested such as information criterions [92, 110, 130] and the inclusion of penalty terms to reduce over fitting [29, 110, 120]. Some authors have integrated prior biological information to aide the reconstruction process [43, 94].

3.3.2 Multi-objective

The number of objectives used is also an area of interest in optimization applications [56, 114], such as minimising error and increasing sparsity, yielding a multi-objective real-world problem. Furthermore, several combinations of these objectives can be used, for example error between the model and the data, sparsity, robustness, connectivity density [90], modularity, biological plausibility [37], etc.

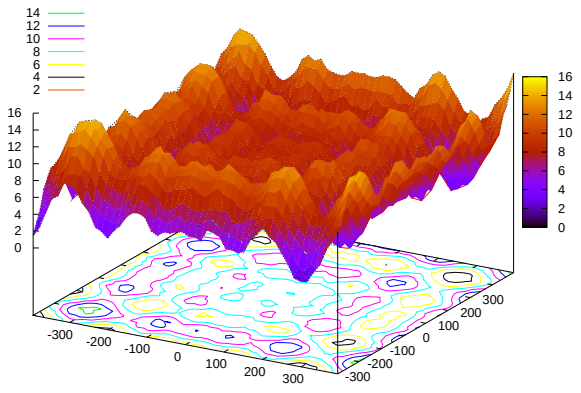

Fig. 4: Complex fitness landscape based on the Schwefel function [127] containing many local minima.

Thus the use of multiple objective functions can be used to infer an accurate network model based on the data, while also using another objective to maintain biological plausibility, as they are able to deal with more complex models [56]. Multiobjective optimization also provides several potential solutions taken from the Pareto front and can be compared and selected by the user based on some preference. Several multi-objective optimization algorithms are available, such as GAs for binary [139] and real valued problems [34], Predator-Prey ES [88], and Pareto-Achived ES [81], all of which are detailed in [33].

3.3.3 Many-objectives

With increasing number of objectives, current optimization algorithms are severely hindered. If there are more than 3 objectives to optimization the problem is known as a many-objective problem, an increasing area of research in the field of optimization [64]. It is possible to use several objectives simultaneously to reconstruct a GRN, as mentioned in Section 3.3.2, to develop a more realistic model. Several algorithms have been proposed to deal with more than three objectives, such as MOPSO [6], NSGA-III [65], GDE3 [83], IBEA [167], and Borg (a framework) [48].

3.4 Multiobjectivization

With all the combinations of objective functions mentioned above it may be possible to utilize the observed phenomenon of multiobjectivization [21, 50, 97] which can aide the optimization search by increasing convergence speed and obtaining global optimum [50, 51]. Multiobjectivization can be achieved by decomposing a single objective in to multiple objectives with similar goals [20, 82], or through the use of additional 'helper' objectives [20, 66] and may also provide more non-dominated solutions with no extra cost to functional evaluation [51]. Multiobjectivization has already been used in GRN parameter inference by Hohm and Zitzler [57], where the authors observed better exploration of the parameter space when using multiple objectives compared with single objectives. However multiobjectivization can also hinder the evolutionary search and performance may be problem specific [20, 51]. Moreover, it has also been observed that equivalent objective functions in different domains can aid evolutionary convergence or hinder the search, indicating that the representation of the objective is also important [146]. Multiobjectivization has been shown to speedup convergence times of inference algorithms, however the practicality of this application to larger networks still requires investigation [132].

3.5 Innovization

Innovization [35] is a process whereby innovative ideas can be found through optimization by analysing the Pareto optimal solutions and observing their special features and commonalities. It is possible to use this technique to discover new and interesting properties and laws of a system

through optimization [9–11]. Analysis of Pareto optimal solutions has previously been used to identify and examine the properties of interesting trade-off solutions [68]. In [56] the authors state that multi-objective optimization can lead to the the discovery of patterns in an organisms structure so is an example of innovization in the reconstruction of GRNs. This parallels the area of big data where analysis of large data sets can provide novel information and may lead to the discovery of new laws or principles. The common example of this is Kepler's Law of planetary motion which was discovered through analysis of Tycho Brahe's systematic astronomical observations, i.e. data analysis [55]. Systems biologists can use multi-objective optimization and innovation to provide novel insights into the structure and characteristics of GRNs using large scale data analytics. Here we can see that reconstruction of GRNs provides a unique interface between optimization techniques and data science.

3.6 Hybrid Algorithms

Due to the large search space of optimizing a model's topology and parameters together, several hybrid algorithms have been suggested to separate these optimization problems. Splitting these process greatly reduce the dimensionality of the problem [7] and can lead to improved algorithm performance and fitness values [92]. Current hybrid algorithms include an artificial neural network and a GA [74], a GP-PSO hybrid [23, 24], a memetic algorithm consisting of a GA and an ES [138], and differential evolution with a local search [110]. In these algorithms the first stage determines the topology of the network and the second determines the parameters of the system. In a similar methodology, nested optimization is also used to separate structure and parameter optimization [134].

4 Big Data in Biology

High-throughput experiments can collect vast amounts of data on gene expression as well as information about the specific techniques and experimental conditions [112]. The cohesion of which can help improve the reconstruction of networks and enable the development of more realistic biological models [134]. Each experiment can produce gene expression levels for thousands of genes at a given time after some biological event giving a genome-wide view of gene expression for the first time [49]. Although currently microarray experiments are expensive and noisy [101], with improvements in technology and process, the constraint of cost will decreases and the number of time steps will increase. Many such experiments use biological replicates, where identical strains of an organism are grown alongside each other under the same conditions to reduce experimental noise. We will soon therefore, have many time points for several biological replicates for each of the genes in an organism for the specific growth conditions, i.e homogeneous data. However, with growth conditions determining the properties of the organism, differing growth conditions are likely to be measured leading to many different heterogeneous data sets. Therefore our ability to combine data sets whilst reducing problems such as heterogeneous noise have to improve [134].

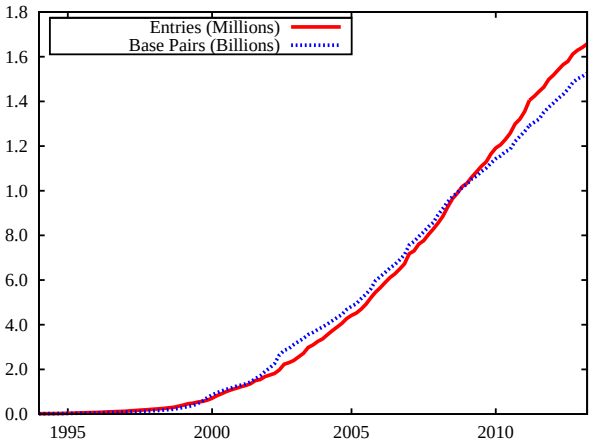

Fig. 5: Number of entries to GenBank database and base pairs contained, which doubles every 18 months [44]. Note the solid line is in units of millions and the dotted line is in units of billions.

Analysing such large quantities of gene expression data is not trivial. This, coupled with their high dimensionally and noise levels [134], biology is rapidly becoming quantitative [12], data intensive science. In order to reconstruct biological networks of significant size and complexity we will need to be able to store, use, share and analyse data efficiently. The so-called curation stage will become increasingly important as high-throughput experiments produce massive quantities of data with varying provenance. At present there is not enough detailed information to accompany the huge amounts of data being collected [49], however, as more effort is made in data curation we will be better able to cope with the data deluge. If we are able to record vast quantities of data, along with relevent information, i.e. provenance, in a usable and accessible way, systems biology can not only predict network structures, but also determine between competing models and develop our understanding of the underlying process of the complex system [155].

Many areas of small scale biology are currently going through a data deluge. The database GenBank which is doubling its number of entries roughly every 18 months [44] and the number of entries and base pairs is shown in Fig. 5. The European Bioinformatics Institution's database of genetic sequencing is also experiencing an exponential increase of entries, doubling at a rate of less than a year [100]. Integration techniques will allow us to combine vast amounts of information from different areas of biology and microbiology and may lead to a unified model of biological systems.

4.1 Big Data and Gene Expression

Many functions of biological systems are unknown and the key to understanding them lies in the deluge of data

being produced [53]. Gene expression data is a practical example of big data in a real world setting. It is obvious how next generation sequencing techniques, such as high-throughput experiments, lead to increases in the volume of data. Beyond more time course data, there is the possibility of replicates, differences in experimental conditions, as well as other organisms in the Genus when comparing similarities across Species. What may be less clear is how expression data, and thus the inference of GRNs, fit with the other 'Four V's of Big Data'.

4.1.1 Veracity

How accurate, precise or reliable the data is, known as data veracity, will depend on the techniques used to conduct the experiments and will undoubtedly improve with the techniques themselves. Biological replicates are used to reduce experimental noise [162] and remove fluctuations from the noisy techniques such as DNA arrays [74]. This in effect is improving the veracity of the data and it is possible to combine data from different high-throughput sources to increase confidence in the data used.

4.1.2 Variety

Aside from obviously different types of data such as steady-state and dynamic data, gene expression experiments rarely have the same growth conditions and thus produce heterogeneous data sets. This provides additional data, increasing data volume, which can be used to test and compare models and may help determine biologically plausible models from those that overfit the data. Data variety is also significantly increased through the use of deletion data, where a gene is deleted in order to determine its regulatory targets [3, 160] and are further examples of heterogeneous data sets that can be integrated in order to learn more about an organism's GRN structure [160]. Another variety of data is from perturbation experiments which are useful in reconstruction of GRNs [42] as the given detailed information on more complex or weaker interactions [160].

4.1.3 Velocity

This aspect of gene expression data comes from two main areas, measurement and analysis. The measurement side includes the experimentation, where high-throughput experiments are occurring in many laboratories leading to a continual increase in the data becoming available. However, the main aspect of the data velocity comes from the analysis of this data, which is still a developing research area. The analysis ranges from turning the raw data into gene expression profiles to using these profiles to infer network structure. The latter is still in its infancy, but in the future we may be able to combine reconstruction algorithms with the experimentation to determine the structure of an organism in real time alongside the initial data analysis. Further we can look at how models built from the data being collected compared to those built from available data to test the impact of the data being collected on model inference.

4.1.4 Volume

Data available on gene expression is increasing at an exponential rate [156]. Beyond biological replicates, repeated experiments, independent verification and simply more time points in the measurements add to increasing volumes of data as discussed previously in Section 4.1.

4.1.5 Linking the Four V's

There is much overlap between the Four V's of big data indicated in the above paragraphs, and changes in one of them can lead to changes in at least one of the others. This implies that big data is more than just a large amount of data points and that several, if not all, of the aspects of big data are linked. It is possible to think of big data as a 3D surface of volume, velocity and variety, as in Fig. 2, with each point on the surface containing an element of veracity. Practically, the link between the Four V's in terms of gene expression is illustrated in Fig. 6. Here we can see that increasing the variety, e.g. conditions, or the veracity, e.g. replicates, increases the volume of data. Moreover, the rate of increase along any of these axis is associated with the velocity of the data, thus there is a strong link between the Four V's of gene expression data.

4.2 Data Integration

Data integration itself is an important issue in biology and is generating considerable interest [49, 55, 86], particularly for heterogeneous data [131, 134]. Heterogenous data is ubiquitous in nature due to the complexity of biology and the lack of standardised experimental protocol resulting in incompatible sources from different data formats [164, pp. 49]. The benefit of models based on multiple data sources was demonstrated in [133], where models were less prone to overfitting and more robust to noise and parameter perturbation. Despite this leading to an increase in the computational complexity, it does help alleviate the problem of underdeterminism [134].

Data integration is vital in data science and the reconstruction of GRNs is already beginning to combine different types of data in order to build more reliable models, such as deletion, perturbation [160], dynamic and steady-state data [166]. Information on biological sparsity [118], the number of network regulators [90], and a shallow architecture [5], can be used as additional objectives or as system constraints to help the optimization of the system. Constraints on the system can help reduce the search space to only biologically plausible regions during the optimization and the additional objectives can potentially be used to increase the algorithm's convergence via multiobjectivization as mentioned in Section 3.4.

Data provenance is also an important aspect of big data and is necessary to reverse engineer biologically plausible networks, particularly when integrating data sets and assessing data veracity. More generally, for different species containing similar network structures and regulators [96], one can integrate data on similar organisms to generalise certain fundamental biological processes.

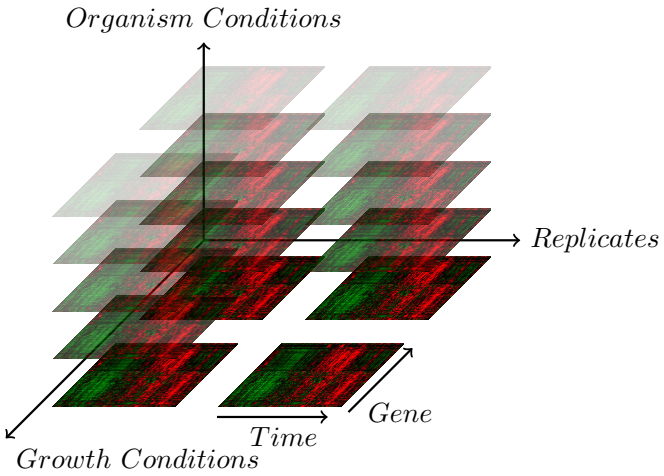

Fig. 6: How mircorray data is big data. Replicates provide homogeneous data sets which can help reduce the veracity of the data. Heterogenous data sets are provided by varying the growth conditions of organism, e.g. the nutrients, and also by varying the organism conditions via perturbation and knockout experiments. These heterogeneous data sets are examples of data variety (see text for details). An increase along any axis leads to an increase in data volume and the rate of increase along any axis is related to the data velocity indicating a link between the elements of the Four V's.

These elements of gene expression data highlight the link from GRNs to big data and the need to incorporate large scale data analytics and curation in order to deal with the 'flood' of data [150]. The curation and integration of data is a growing challenge with the rapid development of many sequencing technologies [164, pp. 51]. The comments on variety alone clearly illustrate the need for improvements in data integration, curation and analysis. Biology is in the midst of a data deluge, gene expression data is a clear example of this and Fig. 6 shows how the different types of experiments from gene expression measurements are leading to changes in data volume, variety and veracity.

4.3 Metaheuristics and Big Data

Metaheuristics are suited for dealing with big data, as they can cope with multiple objectives and constraints based on data provenance or heterogeneous sources. They are also able to deal with large search spaces which are likely to be the case with big data problems, such as inference of GRNs. Metaheuristics are flexible with many kinds of algorithms and frameworks providing scope for tailoring algorithms that are better suited for certain kinds of problems. Developments in the area of so called 'hyper-heuristics' provide a platform for evaluating different metaheuristics for a certain problem.

More specifically, big data problems require researchers to improve data analytics and infer laws or principles from data, similar to Kepler and the laws of planetary motion. A type of metaheuristic, genetic programming, have already been able to infer some of the laws of physics from experimental data [125] and so provide direct evidence that metaheuristics can be used as analytical tools in data problems. Further evidence is provided in Section 3.5, where multi-objective optimization can lead to innovative ideas through innovization. These innovations can discover special features and commonalities between Pareto optimal solutions which may be of great interest to engineering and design disciplines [35].

4.3.1 Optimization Objectives and Data Variety

As discussed in Section 3.3, the type and number of objectives in the reconstruction of GRN from microarray data serves as an interesting problem for the field of optimization. This aspect of reverse engineering GRNs with metaheuristics also ties in with big data, particularly in the data variety. As discussed in Section 4.1.2 and illustrated in Fig. 6, data variety is an important element of gene expression data and therefore must be considered when reconstructing GRNs from data. Beyond comparing inferred networks from different data types for the same organism, which can serve as validation of the network or model, the different data types can be used in the optimization process. A common technique when inferring a GRNs using a metaheuristics is with the objective to minimize the error between the model and the data, recall Section 3.3.1. Here one can have the same objective of error minimization for different data types, such as wild-type and mutant strain gene expression, yielding a multi-objective problem. The use of this heterogeneous data will increase the model complexity as it will have to account for the obvious structural differences between wild-type and mutant/knock-out networks. Despite this, data integration could potentially lead to more robust models by reducing the influence of experimental noise and bias. Furthermore, as high-throughput and many objective optimization techniques improve, we can build networks from several data types, including wild-type, knockout and perturbation data, to help construct more biologically realistic models.

4.3.2 Large Scale Optimization Problems and Data Volume

The problem of high dimensional, large scale optimization problems are currently an area of interest. In 2008 the Congress on Evolutionary Computation (CEC) [142] issued a collection of large scale optimization problems with some strict criteria for anyone taking up the challenge. With 100, 500 and 1000 dimensions in each of the 7 problems, this served as a universal comparison for several large scale optimization algorithms. The performance of some algorithms on these test problems is summarized in Table 1. The performance on such high dimensional problems is highly varied, with at least one order of magnitude difference between the highest and lowest objective value and more than seven orders of magnitude in all but 2 objectives, f_2 and f_7. Even with such challenges, researchers are increasing the dimensionality of problems, e.g. in [95] the authors went up to 2000 dimensions for some of the problems listed in [142]. GRNs,

Table 1: Large Scale Optimization Problems with 1000 Dimensions [142]

Algorithm	Ref	f_1	f_2	f_3	f_4	f_5	f_6	f_7
CCPSO2	[95]	5.18×10^{-13}	7.82×10^{1}	1.33×10^{3}	1.99×10^{-1}	1.18×10^{-3}	1.02×10^{-12}	-1.43×10^{4}
EPUS-PSO	[60]	5.53×10^{2}	4.66×10^{1}	8.37×10^{5}	7.58×10^{3}	5.89×10^{0}	1.89×10^{1}	-6.62×10^{3}
DMS-PSO	[165]	0.00×10^{0}	9.15×10^{1}	8.98×10^{9}	3.84×10^{3}	0.00×10^{0}	7.76×10^{0}	-7.51×10^{3}
MLCC	[158]	8.46×10^{-13}	1.09×10^{2}	1.80×10^{3}	1.37×10^{-10}	4.18×10^{-13}	1.06×10^{-12}	-1.47×10^{4}
sep-CMA-ES	[119]	7.81×10^{-15}	3.65×10^{2}	9.10×10^{2}	5.31×10^{3}	3.94×10^{-4}	2.15×10^{1}	-1.25×10^{4}

which are high dimensional problems are suitable benchmarks for these kind of algorithms. Analysing genome-wide high-throughput data for even a small organism of a few thousand genes will have an enormous dimensionality. At present the best solution for the underdeterminism problem is to produce more time series data and more replicates, which will leave us with a high dimensional problem with vast amounts of data, i.e. a big data and large scale optimization problem.

4.3.3 Uncertain Optimization, Data Veracity and Data Velocity

Uncertainty in optimization problems can be divided in to four categories: noisy fitness functions, post-optimization design or environmental parameter perturbation, the use of approximate fitness functions and changes to the global optimum over time [67]. The area of evolutionary dynamic optimization has received much attention in the last 20 years due to its application to real world problems [108]. This area of research is important to GRN reconstruction as there is much uncertainty, most notably the form of the objective function. The links to data veracity here are clear, with uncertainty in the data there is also uncertainty in the optimization. An argument may be made for data uncertainties propagating through and leading to uncertainty in the fitness function, particularly if the objective is minimizing the error between the model and the data. However, for any objective function, this also falls into the class of robustness, i.e. how robust is the GRN to noise in the data? We have already discussed the noise in high-throughput data in Section 4 and how this links to data veracity in Section 4.1.1, further showing the overlap between big data and optimization for GRN reconstruction.

A special class of dynamic problems are dynamic optimization problems (DOPs) which require online optimization [108]. In Section 4.1.3 it was mentioned that in the future we can combine experimentation and modelling to reverse engineer GRNs while data is being recorded. This would enable researchers to observe how growth conditions and the environmental can cause a change in GRNs, i.e. if a certain environmental perturbation causes a biological switch or activates a previously unknown pathway. To cope with DOPs may require significant computational power to perform the optimization online, i.e. in real time, thus relating to data velocity. However despite these computational challenges, our understanding of biological structures and interactions could dramatically improve.

5 GRNs in Artificial Life

ALife studies using GRNs have investigated the simple addition and extension of a system by introducing new mechanisms [59]. Here, in [59], the author showed that a simple model, with an abstraction level at the cell, was able to model developmental growth and regeneration distinctively. Regeneration was also studied using a similar model in [126], where it was observed that high morphological plasticity can result in the regenerative ability of artificial organisms. The authors also note that the high morphological plasticity results from stable growth and the controlled removal of cells, but at a high energy cost.

Through computational simulations von Dassow et al. [31] were able to show the extent of the robustness of the Segment Polarity in *Drosophila*, as discussed in Section 2.3. Computational investigations into the properties of GRNs can improve our fundamental understanding of different types of networks, such as scale-free networks that are more robust than randomly connected networks [2], however are sensitive to perturbations to the hub genes [78]. Redundancy, which has been shown to improve evolvabillity, is also linked to robustness, where simplified GRNs are less stable against mutations than those with redundancies [126].

The study of more fundamental properties of GRNs are vital, such as how we can model several incoming iterations to one node or gene. This is essential and related to the complexity of biological systems, observed network redundancies, hub node topologies and motifs structures. For one gene to regulate another it must bind to the *cis*-region of the target, which may have several *cis*-region each of which may be for a different regulator [80]. The necessity for combining biological interactions is highlighted by the feedforward loop (FFL) motif [129]. The concentration of this ubiquitous structure is independent of subnetwork size in *E. coli* and appears significantly more often than in a random network [103]. It is often assumed that regulation from multiple sources can be modelled using logic gates such as 'AND' and 'OR' operations [5]. However, Schilstra and Nevhaniv [124] investigated the role of combinatorial logic in gene expression for a target with two regulators in several regulator set-ups. The authors observed that although these set-ups 'mimic' Boolean logic, the binding of the regulators can not be combined (or decomposed) in the same way. The only exception of this is the case where the two regulators bind independently, i.e. at different *cis*-regions of the target. For all other cases (both conjunctive and disjunctive) competitive, ordered and joint binding at the

cis-regions, the laws of combinatorial Boolean logic are not sufficient to model the systems output.

The topics discussed above highlight the importance of investigation into not only complex network structures and dynamics, but also fundamental principles and concepts of biological networks. Understanding the basics of biological interactions and process is essential for the development of 'soft' ALife, but will ultimately lead to improvements in the 'hard' and 'wet' areas. This fundamental understanding is critical for building biologically plausible models, both the synthesis from basic principals and from reverse engineering from data. Developments in ALife and basic biological principals are essential for investigations into GRNs inference.

6 Current Reconstruction Algorithms

There are many reconstruction algorithms available and several reviews comparing them with different data sets, some of which are detailed in Table 2. In [12] several algorithms were tested against steady-state and dynamic data sets of varying sizes, however none of the algorithms tested were able to reconstruct networks from dynamic time-series data, of any size, significantly better than random. The reconstruction algorithms in [46] used various sized networks and concluded that none of the algorithms tested outperform the others. The authors also concluded that not only were none of them able to reconstruct the true network for all data sets, they all had low precision and resulted in many false positives. In [132] it was noted EA methods are favoured due to the lack of data, however hybrid algorithms are needed for larger networks. The authors also stated that as hybrid methods are computational expensive, parallel implementation is vital. Penfold and Wild [113] found that reconstruction from time course was nearly always better than those from systematic knockout experiments. They also noted that nonlinear dynamical systems (NDS) methods performed the best based on time series data, however they scale with the number of observations and are impractical for cases where many time courses are generated. This becomes particularly problematic when the networks are large, for 100 genes with 21 time points and 10 replicates (totalling 210 observations per gene), NDS methods require 48 hrs of computational time per gene [113], or 200 days for the entire network.

Bayesian, informatics and ODE networks are all reliant on the available data for reconstruction, where moderate size networks 10-20 time points is insufficient for statistical reconstruction [159]. Computationally determined networks may also be considerably smaller than experimentally measured networks, e.g. [27] where only 20 out of 600 genes were reconstructed due to the high dimensionality of the problem. It may also be impossible to distinguish between models derived from small data sets [41]. The restricted network size for Bayesian reconstruction is due to the superexponential search space, though this can be reduced by integrating additional biological knowledge [38]. Thus using Bayesian networks to reverse engineer GRNs will significantly benefit from the larger and more diverse data sets from the data deluge.

Table 2: Current reconstruction algorithms. Model type acronyms are ordinary differential equation (ODE), neural network (NN), dynamic Bayesian network (DBN) and non-linear dynamical systems (NDS)

Algorithm	Model type	Ref.	Reviewed
ARACNE	Relevance network	[99]	[12, 46]
Banjo	Bayesian network	[161]	[12, 46]
NIR/MNI	ODE	[18, 42]	[12]
GNRevealer	Neural network	[47]	[46]
LDST	State space	[117]	[46]
GeneNet	Graphical Gaussian	[123]	[46]
ParCorA	Pearson / Spearman	[32]	[46]
DE+AIC	ODE	[109, 110]	[132]
GA+ANN	ODE + NN	[74]	[132]
GLSDC	ODE	[148]	[132]
PEACE1	ODE	[77]	[132]
GA+ES	ODE	[137]	[132]
G1DBn	DBN	[89]	[113]
VBSSM	DBN	[14]	[113]
TSNI	ODE	[13]	[113]
GP4GRN	NDS	[1]	[113]
CSI	NDS	[79, 113]	[113]
GCCA	Granger causality	[128]	[113]

There is an urgent need for reconstruction tools [92], particularly those that can deal with a large number of genes and can reverse engineer networks based on real data. Many algorithms use artificial data as a benchmarking tool as the true network is known and the performance can be quantified. Although the artificial data can be useful for comparing algorithms and see how models use the data [115], they are not as complex as real biological data [46]. These benchmark tests often use multiple data sets that increase with the number of genes in the target network, however with the expense of current experiments [7, 113] this is not realistic.

Performance of reconstruction algorithms will improve with additional data, however, advances in reverse engineering algorithms is desperately needed to incorporate data integration techniques. This may help reduce the problem of underdeterminism and allow the algorithms to cope with the high levels of noise within mircoarray data sets, as well as any missing data points.

Figure 7 shows how the computational time for network reconstruction increases with increasing network size for several data and model types. Particularly notable is the large runtimes for real data sets and the significant spread of the runtimes for networks of the same size. This spread is likely to be due to the quality of the data available, which will vary significantly between experiments and will not be present in artificially generated data sets. Model complexity is also a factor as this can significantly effect the computational run time based on the number of parameters alone. It is also clear that reconstruction based on dynamic data sets, even for small networks are computationally very expensive. Many authors do not include information about the runtimes of their algorithms so the available information is relatively small. Lines of best fit are included using the gunplot fit function [22] using $y = x^m c$. Note that as only two different network sizes

are available for the artifical topological models, no line of best fit is included. It also should be noted that the four right most real dynamic (square) data points are omitted from the fitting as the authors in [74] only show their topological results, the results coincidently fit well with the other topological models based on real data (circles). It does however illustrate the need for improvements in computational techniques, not only in the reconstruction algorithms and methods, but in the execution of the algorithm and other computational tools [73]. The lines in Fig. 7 indicate that artificial time series data is not as complex as real data sets demonstrated by an increase in average runtime for networks of the same size. There is little research into reconstruction of GRNs based on real data compared to artificial data [49], which are ultimately just benchmarks for algorithms but are not adding to our understanding of biological networks. More effort is needing in modelling real networks, however it should also be stressed that part of the reason for this is the lack of available data sets from real experiments.

7 Computational Improvements

Parallel computing is an obvious solution for the runtimes of algorithms which can be significantly reduced by splitting them over several cores. This does, however, require some expertise and programming at a low level of abstraction [87] making the algorithm more problem-specific and less useful for general reverse engineering of networks. Xiao et al. in [156] used parallel processing to reconstruct a network of 1000 genes based on artificial perturbation data and observed a parallel speedup of 20 times with 32 cores. Although they noticed a sharp decrease in parallel efficiency when using up to four cores, due to information sharing, there is little change between 4 and 32 cores. Even with this continual decrease in efficiency with additional nodes, the authors reduced the computational runtime from 108 *mins* to 5 *mins*.

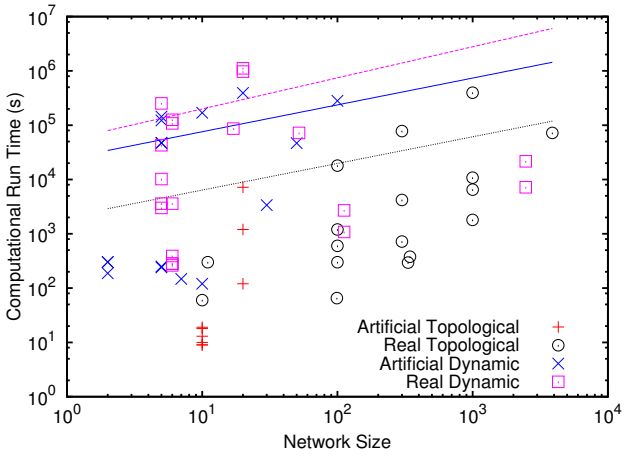

Fig. 7: Reconstructed network size versus computational runtime. Data types are either from artifical networks (A) or from real biological experiments (R) and are either dynamic (D) or topological (T) models as indicated in the figure key. Lines of best fit are included for RT (dotted), AD (solid) and RD (dashed) data, see text for details. Data collected from [12, 57, 72, 74, 98, 102, 105, 107, 109, 110, 123, 132, 149, 156, 157, 160, 161, 166].

This speedup can be furthered using graphic processing units (GPUs), which can contain 1000s of cores and reduce a runtime of days into minutes, though this requires even more expertise than standard parallel computing. Despite the barrier of knowledge, parallel computing is becoming a necessity [132] for network reconstruction, and as the quantity and quality of data increases so must our computational abilities. Some reconstruction methods already take hours to reconstruct small networks [77, 107, 110] and others are already running on computer clusters [160]. Methods that are not intrinsically parallel, such as the S-System Eq. (3), can be decoupled so that each connection can be treated seperately allowing parallelization. Such techniques can significantly reduce the problem dimension size and search space [74, 110, 134, 151].

Further computational improvements can be made at the algorithm level. As available data increases and data integration techniques improve and become more common, statistical based reconstruction methods will benefit from increasing levels of usable data. However how the algorithms deal with data and the rate at which information is processed and then used in the reconstruction needs to improve so that large and more complex networks can be reconstructed. Improvements in efficiency in algorithm operations, such as solution evaluation and selection may lead to significant speedups.

Although the problem of underdeterminism of GRNs is addressed with increasing the amount of data, metaheuristics are still an attractive option due to their flexibility and search power. Even in an ideal case where high quality time series data is plentiful, there may still be a considerable search space for connection parameters, and even for the form of connection between genes. Not only can metaheuristics be used in connection topology and parameterization of a network, they also offer a platform for optimizing biological objectives whilst also being intrinsically parallel [134]. Several objective functions can be compared and used together to investigate their effects on the final network which may help to identify important biological objectives in addition to the potential benefits of multiobjectivization. Furthermore, with the development of many-objective optimization, there is potential for using several objectives to help steer the search towards biologically plausible solutions.

In [151] the authors noted that 95% of the optimization time for larger networks was spent on numerical integration. This can often lead to divergence or even algorithm failure [150]. However this does indicate the potential for significant improvements to runtime efficiency providing improvements to the mathematical and computational techniques occur, e.g. parallelizing integration calculations for newtork reconstruction.

8 Conclusions

In this paper we have introduced the problem of the reverse engineering of GRNs from expression data and how

this links optimization and big data. We overview some common GRN models and discuss the difference of topological and parameterized models and the importance of combining the two. Some metaheuristic methods are over viewed and several possible objective functions are suggested for the reverse engineering for both single and multi-objective optimization. The use of additional objectives can not only aid optimization convergence but can also provide biologically plausible solutions. Moreover we argue the reverse engineering of GRNs is a real-world platform for many objective optimization. Other areas of optimization, such as the phenomenon of multiobjectivization, can be investigated for reconstructing GRNs by comparing algorithm convergence from single and multi-objective set ups.

We go beyond the inference of GRNs and link optimization to data science directly through innovization [35]. Here solutions to a multi-objective optimization problem can be used to discover laws and properties of a system, and therefore is a form of data science. We discuss the data deluge in biology and detail how gene expression data relates to each of the component areas of big data; veracity, variety, velocity and volume. Further, we argue that these areas of big data are strongly linked for gene expression data and illustrate this in Fig. 6. We examine the importance of data integration due to the heterogeneity, or the variety, of biological data and how this can be used with optimization algorithms. The synergy between big data and optimization is extended by linking the components of big data to specific areas of optimization research, and generalises this beyond GRN inference. Specifically we link the number of objectives with data variety, large scale optimization with data volume and uncertain optimization with data veracity and velocity. We argue that the reconstruction of GRNs from gene expression data provides a data intensive problem that applies to many areas of optimization research.

We state the importance of ALife in the progression of GRN inference. The study of ALife using GRNs is well established and some key areas are detailed in Section 5. Here we note that our understanding of the properties of large networks, as well as the fundamental biological process studied in ALife, will improve GRN reconstruction alongside the data deluge. Although not directly related to big data, ALife is vital for the fundamentals of the interactions in GRNs. Furthermore, ALife can aide inference by providing evidence for possible biological 'objectives'

Some current reconstruction algorithms are over viewed and the limitations of these methods are discussed. We address the issue of computational runtime with increasing network size for many current algorithms and note the significant difference between real and synthetic time course data. This leads to a problem when using artificial benchmark data sets to compare algorithms as they do not reflect real biological systems. Because of this issue we also investigate the use of parallelization in reconstruction algorithms and argue its necessity in the future for practical applications to real time course data.

Reconstruction algorithms must be able to scale up to tens of thousands of genes [147] in order to model higher level organisms. The future of reconstruction algorithms is likely to contain a combination of data intensive and optimization processes in order to analyse the data required for the optimization of a large network's search space. This big data optimization synergy requires both biologists and computer scientists to work together and share not only data, but also expertise, knowledge and processes.

Acknowledgements This work was funded by an Engineering and Physics Science Research Council (EPSRC) Doctoral Trainning Centre (DTC) studentship at the University of Surrey.

References

1. Äijö, T., Lähdesmäki, H.: Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. Bioinformatics **25**(22), 2937–2944 (2009)
2. Albert, R., Jeong, H., Barabasi, A.L.: Error and attack tolerance of complex networks. Nature **406** (2000)
3. Allenby, N.E.E., Laing, E., Bucca, G., Kierzek, A.M., Smith, C.P.: Diverse control of metabolism and other cellular processes in streptomyces coelicolor by the phop transcription factor: genome-wide identification of in vivo targets. Nucleic Acids Research (2012)
4. Alon, U.: An Introduction to Systems Biology: Design Principles of Biological Circuits. Chapman & Hall/CRC (2006)
5. Alon, U.: Network motifs: theory and experimental approaches. Nat. Rev. Genet. **8**, 450 – 461 (2007)
6. Alvarez-benitez, J.E., Everson, R.M., Fieldsend, J.E.: A mopso algorithm based exclusively on pareto dominance concepts. In: In Third International Conference on Evolutionary MultiCriterion Optimization, EMO 2005, pp. 459–473. SpringerVerlag (2005)
7. Ando, S., Iba, H.: Inference of gene regulatory model by genetic algorithms. In: Evolutionary Computation, 2001. Proceedings of the 2001 Congress on, vol. 1, pp. 712 –719 (2001)
8. Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., Teichmann, S.A.: Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology **14**(3), 283 – 291 (2004)
9. Bandaru, S., Deb, K.: Automated innovization for simultaneous discovery of multiple rules in bi-objective problems. In: R. Takahashi, K. Deb, E. Wanner, S. Greco (eds.) Evolutionary Multi-Criterion Optimization, *Lecture Notes in Computer Science*, vol. 6576, pp. 1–15. Springer Berlin Heidelberg (2011)
10. Bandaru, S., Deb, K.: Towards automating the discovery of certain innovative design principles through a clustering-based optimization technique. Engineering Optimization **43**(9), 911–941 (2011)

11. Bandaru, S., Deb, K.: A dimensionally-aware genetic programming architecture for automated innovization. In: R. Purshouse, P. Fleming, C. Fonseca, S. Greco, J. Shaw (eds.) Evolutionary Multi-Criterion Optimization, *Lecture Notes in Computer Science*, vol. 7811, pp. 513–527. Springer Berlin Heidelberg (2013)
12. Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D.: How to infer gene networks from expression profiles. Mol Syst Biol **3** (2007)
13. Bansal, M., Gatta, G.D., di Bernardo, D.: Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics **22**(7), 815–822 (2006)
14. Beal, M.J., Falciani, F., Ghahramani, Z., Rangel, C., Wild, D.L.: A bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics **21**(3), 349–356 (2005)
15. Bedau, M.A.: Artificial life: organization, adaptation and complexity from the bottom up. Trends in Cognitive Sciences **7**(11), 505 – 512 (2003)
16. Bell, G.: The Fourth Paradigm: Data-Intensive Scientific Discovery, 1st edn., chap. Foreword, pp. xi–xv. Microscoft Research (2009)
17. Ben-Gal, I.: Bayesian Networks. John Wiley & Sons, Ltd (2008)
18. di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E., Collins, J.J.: Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nature Biotechnology (3), 377383 (2005)
19. Beyer, H.G., Schwefel, H.P.: Evolution strategies a comprehensive introduction. Natural Computing **1**(1), 3–52 (2002)
20. Brockhoff, D., Friedrich, T., Hebbinghaus, N., Klein, C., Neumann, F., Zitzler, E.: Do additional objectives make a problem harder? In: Proceedings of the 9th annual conference on Genetic and evolutionary computation, GECCO '07, pp. 765–772. ACM, New York, NY, USA (2007)
21. Brockhoff, D., Friedrich, T., Hebbinghaus, N., Klein, C., Neumann, F., Zitzler, E.: On the effects of adding objectives to plateau functions. Evolutionary Computation, IEEE Transactions on **13**(3), 591 –603 (2009)
22. Bröker, H.B., et al.: gnuplot 4.6: An interactive plotting program (2012). URL http://www.gnuplot.info/
23. Cai, X.: A multi-objective gp-pso hybrid algorithm for gene regulatroy network modeling. Ph.D. thesis, Kansas State University, Manhatten, Kansas (2009)
24. Cai, X., Koduru, P., Das, S., Welch, S.M.: Simultaneous structure discovery and parameter estimation in gene networks using a multi-objective gp-pso hybrid approach. Int. J. Bioinformatics Res. Appl. **5**(3), 254–268 (2009)
25. Chen, B.S., Hsu, C.Y., Liou, J.J.: Robust design of biological circuits: Evolutionary systems biology approach. Journal of Biomedicine and Biotechnology **2011**, 14 (2011)
26. Chen, L.: Computational systems biology on networks and dynamics. In: Optimization and Systems Biology, pp. 5–12 (2007)
27. Chen, X.w., Anantha, G., Wang, X.: An effective structure learning method for constructing gene networks. Bioinformatics **22**(11), 1367–1374 (2006)
28. Chiquet, J., Grandvalet, Y., Ambroise, C.: Inferring multiple graphical structures. Statistics and Computing **21**(4), 537–553 (2011)
29. Chowdhury, A., Chetty, M., Vinh, X.: On the reconstruction of genetic network from partial microarray data. In: T. Huang, Z. Zeng, C. Li, C. Leung (eds.) Neural Information Processing, *Lecture Notes in Computer Science*, vol. 7663, pp. 689–696. Springer Berlin Heidelberg (2012)
30. Crombach, A., Hogeweg, P.: Evolution of evolvability in gene regulatory networks. PLoS Comput Biol **4**(7), e1000,112 (2008)
31. von Dassow, G., Meir, E., Munro, E.M., Odell, G.M.: The segment polarity network is a robust developmental module. Nature **406**, 188–192 (2000). DOI 10.1038/35018085. URL http://dx.doi.org/10.1038/35018085
32. De La Fuente, A., Bing, N., Hoeschele, I., Mendes, P.: Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics **20**(18), 3565–3574 (2004)
33. Deb, K.: Multi-Objective Optimisation using Evolutionary Algortihms, 1st edn. Wiley, Kanpur, India (2001)
34. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: Nsga-ii. Evolutionary Computation, IEEE Transactions on **6**(2), 182 –197 (2002)
35. Deb, K., Srinivasan, A.: Innovization: innovating design principles through optimization. In: Proceedings of the 8th annual conference on Genetic and evolutionary computation, GECCO '06, pp. 1629–1636. ACM, New York, NY, USA (2006)
36. DSouza, R.G.L., Sekaran, K.C., , Kandasamy, A.: A phenomic algorithm for reconstruction of gene networks. World Academy of Science, Engineering and Technology **31**, 53–58 (2007)
37. D'Souza, R.G.L., Sekaran, K.C., A, K.: Reconstruction of gene networks using phenomic algorithms. International Journal of Artificial Intelligence & Applications **1**(2) (2010)
38. Filkov, V.: Identifying Gene Regulatory Networks from Gene Expression Data, chap. 27, pp. 27–1–27–29. Chapman and Hall/CRC (2005)
39. Fioravanti, F., Helmer-Citterich, M., Nardelli, E.: Modeling gene regulatory network motifs using statecharts. BMC Bioinformatics **13**(4), 1–12 (2012)
40. Frank, K., Rckl, M., Nadales, M.J.V., Robertson, P., Pfeifer, T.: Comparison of exact static and dynamic bayesian context inference methods for activity recognition. In: PerCom Workshops, pp. 189–195. IEEE (2010)

41. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. Journal of Computational Biology **7**, 601–620 (2000)
42. Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J.: Inferring genetic networks and identifying compound mode of action via expression profiling. Science **301**(5629), 102–105 (2003)
43. Geier, F., Timmer, J., Fleck, C.: Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. BMC Systems Biology **1** (2007)
44. GenBank: National center for biotechnology information, genetic sequence data bank (June 15 2013). URL ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt
45. Gonze, D.: Coupling oscillations and switches in genetic networks. Biosystems **99**(1), 60 – 69 (2010)
46. Hache, H., Lehrach, H., Herwig, R.: Reverse engineering of gene regulatory networks: a comparative study. EURASIP J. Bioinformatics Syst. Biol. **2009**, 8:1–8:12 (2009)
47. Hache, H., Wierling, C., Lehrach, H., Herwig, R.: Reconstruction and validation of gene regulatory networks with neural networks. In: The 2nd Foundations of Systems Biology in Engineering Conference, FOSBE 2007, pp. 319–324 (2007)
48. Hadka, D., Reed, P.: Borg: An auto-adaptive many-objective evolutionary computing framework. Evolutionary Computation **21**, 231–259 (2013)
49. Hallinan, J.: Gene Networks and Evolutionary Computation, pp. 67–96. John Wiley & Sons, Inc. (2007)
50. Handl, J., Kell, D.B., Knowles, J.: Multiobjective optimization in bioinformatics and computational biology. IEEE/ACM Trans. Comput. Biol. Bioinformatics **4**(2), 279–292 (2007)
51. Handl, J., Lovell, S.C., Knowles, J.: Investigations into the effect of multiobjectivization in protein structure prediction. In: Proceedings of the 10th international conference on Parallel Problem Solving from Nature: PPSN X, pp. 702–711. Springer-Verlag, Berlin, Heidelberg (2008)
52. Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W.: From molecular to modular cell biology. Nature **402**, C47–C52 (1999)
53. Haynes, W.A., Higdon, R., Stanberry, L., Collins, D., Kolker, E.: Differential expression analysis for pathways. PLoS Comput Biol **9**(3), e1002,967 (2013)
54. Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., Guthke, R.: Gene regulatory network inference: Data integration in dynamic modelsa review. Biosystems **96**(1), 86 – 103 (2009)
55. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery, 1st edn. Microscoft Research (2009)
56. Higuera, C., Villaverde, A.F., Banga, J.R., Ross, J., Morn, F.: Multi-criteria optimization of regulation in metabolic networks. PLoS ONE **7**(7), e41,122 (2012)
57. Hohm, T., Zitzler, E.: Multiobjectivization for parameter estimation: a case-study on the segment polarity network of drosophila. In: F. Rothlauf, et al. (eds.) GECCO '09: Genetic and Evolutionary Computation Conference (GECCO 2009), pp. 209–216. ACM, New York, NY, USA (2009)
58. Hoon, M.D., Imoto, S., Miyano, S.: Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations. In: Pac. Symp. Biocomput, pp. 17–28 (2003)
59. Hotz, P.E.: Exploring regenerative mechanisms found in flatworms by artificial evolutionary techniques using genetic regulatory networks. In: Evolutionary Computation, 2003. CEC '03. The 2003 Congress on, vol. 3, pp. 2026–2033 (2003)
60. Hsieh, S.T., Sun, T.Y., Liu, C.C., Tsai, S.J.: Solving large scale global optimization using improved particle swarm optimizer. In: Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on, pp. 1777–1784 (2008)
61. Iba, H., Mimura, A.: Inference of a gene regulatory network by means of interactive evolutionary computing. Inf. Sci. Inf. Comput. Sci. **145**(3-4), 225–236 (2002)
62. IBM: What is big data? (3/7/13). URL http://www-01.ibm.com/software/data/bigdata/
63. Ingram, P., Stumpf, M., Stark, J.: Network motifs: structure does not determine function. BMC Genomics **7**(1), 1–12 (2006)
64. Ishibuchi, H., Tsukamoto, N., Nojima, Y.: Evolutionary many-objective optimization: A short review. In: Proceedings of Congress on Evolutionary Computation, CEC, pp. 2424–2431 (2008)
65. Jain, H., Deb, K.: An improved adaptive approach for elitist nondominated sorting genetic algorithm for many-objective optimization. In: R. Purshouse, P. Fleming, C. Fonseca, S. Greco, J. Shaw (eds.) Evolutionary Multi-Criterion Optimization, *Lecture Notes in Computer Science*, vol. 7811, pp. 307–321. Springer Berlin Heidelberg (2013)
66. Jensen, M.T.: Helper-objectives: Using multi-objective evolutionary algorithms for single-objective optimisation. Journal of Mathematical Modelling and Algorithms **3**, 323–347 (2004)
67. Jin, Y., Branke, J.: Evolutionary optimization in uncertain environments-a survey. Evolutionary Computation, IEEE Transactions on **9**(3), 303–317 (2005)
68. Jin, Y., Gruna, R., Sendhoff, B.: Pareto analysis of evolutionary and learning systems. Frontiers of Computer Science in China **3**(1), 4–17 (2009)
69. Jin, Y., Meng, Y.: Emergence of robust regulatory motifs from in silico evolution of sustained oscillation. Biosystems **103**(1), 38 – 44 (2011)
70. de Jong, H.: Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology **9**, 67–103 (2002)
71. de Jong, H., Geiselmann, J.: Modeling and simulation of genetic regulatory networks by ordinary differential equations. In: Genomic Signal Processing and Statistics, p. 201239. Hindawi Publishing Cor-

poration, New York (2005)
72. Kabir, M., Noman, N., Iba, H.: Reverse engineering gene regulatory network from microarray data using linear time-variant model. BMC Bioinformatics **11**, S56 (2010)
73. Karlebach, G., Shamir, R.: Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol **9**, 770 – 780 (2008)
74. Keedwell, E., Narayanan, A.: Discovering gene networks with a neural-genetic hybrid. IEEE/ACM Trans. Comput. Biol. Bioinformatics **2**(3), 231–242 (2005)
75. Khammash, M.: Reverse engineering: the architecture of biological networks. BioTechniques **44**, 323,329 (2008)
76. Khammash, M., El-Samad, H.: Systems biology: from physiology to gene regulation. Control Systems, IEEE **24**(4), 62–76 (2004)
77. Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., Tomita, M.: Dynamic modeling of genetic networks using genetic algorithm and s-system. Bioinformatics **19**(5), 643–650 (2003)
78. Kitano, H.: Computational systems biology. Nature **420** (2002)
79. Klemm, S.: Causal structure identication in nonlinear dynamical systems (2008)
80. Knabe, J.F., Wegner, K., Nehaniv, C.L., Schilstra, M.J.: Genetic algorithms and their application to in silico evolution of genetic regulatory networks. In: D. Fenyö (ed.) Computational Biology, *Methods in Molecular Biology*, vol. 673, pp. 297–321. Humana Press (2010)
81. Knowles, J.D., Corne, D.W.: Approximating the nondominated front using the pareto archived evolution strategy. Evol. Comput. **8**(2), 149–172 (2000)
82. Knowles, J.D., Watson, R.A., Corne, D.: Reducing local optima in single-objective problems by multi-objectivization. In: Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization, EMO '01, pp. 269–283. Springer-Verlag, London, UK, UK (2001)
83. Kukkonen, S., Lampinen, J.: Gde3: the third evolution step of generalized differential evolution. In: Evolutionary Computation, 2005. The 2005 IEEE Congress on, vol. 1, pp. 443–450 Vol.1 (2005)
84. Kuo, P.D., Banzhaf, W., Leier, A.: Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. Biosystems **85**(3), 177 – 200 (2006)
85. Kwon, Y.K., Cho, K.H.: Quantitative analysis of robustness and fragility in biological networks based on feedback dynamics. Bioinformatics **24**(7) (2008)
86. Lanely, D.: 3d data management: Controlling data volume, velocity and variety. Tech. rep., Application Delivery strategies: META Group (2001)
87. Larus, J., Gannon, D.: The Fourth Paradigm: Data-Intensive Scientific Discovery, 1st edn., chap. Multicore computing and scientific discovery, pp. 125–129. Microsoft Research (2009)
88. Laumanns, M., Rudolph, G., Schwefel, H.P.: A spatial predator-prey approach to multi-objective optimization: A preliminary study. In: Proceedings of the 5th International Conference on Parallel Problem Solving from Nature, PPSN V, pp. 241–249. Springer-Verlag, London, UK, UK (1998)
89. Lèbre, S.: Inferring dynamic genetic networks with low order independencies. Statistical Applications in Genetics and Molecular Biology **8**, 1–38 (2009)
90. Leclerc, R.D.: Survival of the sparsest: robust gene networks are parsimonious. Molecular Systems Biology pp. 1–6 (2008)
91. Lee, W.P., Hsiao, Y.T.: Inferring gene regulatory networks by incremental evolution and network decomposition. In: Optimization and Systems Biology, pp. 311–324 (2008)
92. Lenser, T., Hinze, T., Ibrahim, B., Dittrich, P.: Towards evolutionary network reconstruction tools for systems biology. In: E. Marchiori, J. Moore, J. Rajapakse (eds.) Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics, *Lecture Notes in Computer Science*, vol. 4447, pp. 132–142. Springer Berlin Heidelberg (2007)
93. Li, C., Chen, L., Aihara, K.: A systems biology perspective on signal processing in genetic network motifs [life sciences]. Signal Processing Magazine, IEEE **24**(2), 136 –147 (2007)
94. Li, J., Zhang, X.S.: An optimization model for gene regulatory network reconstruction with known biological information. In: Optimization and Systems Biology, pp. 35–44 (2007)
95. Li, X., Yao, X.: Cooperatively coevolving particle swarms for large scale optimization. Evolutionary Computation, IEEE Transactions on **16**(2), 210–224 (2012)
96. Liu, Y., Niculescu-Mizil, A., Lozano, A.C., Lu, Y.: Temporal graphical models for cross-species gene regulatory network discovery. J. Bioinformatics and Computational Biology **9**(2), 231–250 (2011)
97. Lochtefeld, D., Ciarallo, F.: Multiobjectivization via helper-objectives with the tunable objectives problem. Evolutionary Computation, IEEE Transactions on **16**(3), 373 –390 (2012)
98. Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S., Eguchi, Y.: Development of a system for the inference of large scale genetic networks. Pac Symp Biocomput. pp. 446–458 (2001)
99. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., Califano, A.: ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics **7**((Suppl 1)), S7 (2006)
100. Marx, V.: Biology: The big challenges of big data. Nature **498**, 255–260 (2013)
101. McLachlan, G.J., Do, K.A., Ambroise, C.: Analyzing Microarray gene Expression Data. Wiley-Interscience (2004)
102. Mendoza, M.R., Bazzan, A.L.C.: Evolving random boolean networks with genetic algorithms for regula-

tory networks reconstruction. In: Proceedings of the 13th annual conference on Genetic and evolutionary computation, GECCO '11, pp. 291–298. ACM, New York, NY, USA (2011)
103. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., net, U.: Network motifs: Simple building blocks of complex networks. Science **298**(5594), 824–827 (2002)
104. Mitchell, M.: An Introduction to Genetic Algorithms. The MIT Press (1999)
105. Mondal, B., Sarkar, A., Hasan, M., Noman, N.: Reconstruction of gene regulatory networks using differential evolution. In: Computer and Information Technology (ICCIT), 2010 13th International Conference on, pp. 440 –445 (2010)
106. Morishita, R., Imade, H., Ono, l., Ono, N., Okamoto, M.: Finding multiple solutions based on an evolutionary algorithm for inference of genetic networks by s-system. In: Evolutionary Computation, 2003. CEC '03. The 2003 Congress on, vol. 1, pp. 615–622 (2003)
107. Nakayama, T., Seno, S., Matsuda, H.: Inference of s-system models of gene regulatory networks using immune algorithm. J Bioinform Comput Biol **9**, 75–86 (2011)
108. Nguyen, T.T., Yang, S., Branke, J.: Evolutionary dynamic optimization: A survey of the state of the art. Swarm and Evolutionary Computation **6**(0), 1 – 24 (2012)
109. Noman, N., Iba, H.: Inference of genetic networks using s-system: information criteria for model selection. In: Proceedings of the 8th annual conference on Genetic and evolutionary computation, GECCO '06, pp. 263–270. ACM, New York, NY, USA (2006)
110. Noman, N., Iba, H.: Inferring gene regulatory networks using differential evolution with local search heuristics. IEEE/ACM Trans. Comput. Biol. Bioinformatics **4**(4), 634–647 (2007)
111. Noor, A., Serpedin, E., Nounou, M., Nounou, H., Mohamed, N., Chouchane, L.: Information theoretic methods for modeling of gene regulatory networks. In: Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on, pp. 418–423 (2012)
112. Pastrello, C., Otasek, D., Fortney, K., Agapito, G., Cannataro, M., Shirdel, E., Jurisica, I.: Visual data mining of biological networks: One size does not fit all. PLoS Comput Biol **9**(1), e1002,833 (2013)
113. Penfold, C.A., Wild, D.L.: How to infer gene networks from expression profiles, revisited. Interface Focus (2011)
114. Purshouse, R.C., Fleming, P.J., Fonseca, C.M., Greco, S., Shaw, J.: 7th international conference, emo 2013, sheffield, uk, march 19-22, 2013. proceedings. In: Evolutionary Multi-Criterion Optimization, vol. 7811. Springer Berlin Heidelberg (2013)
115. Quackenbush, J.: Computational analysis of microarray data. Nature Reviews Genetics (6), 418–427 (2001)
116. Ramons, A.F., Innocentini, G., Forger, F.M., Hornos, J.E.: Symmetry in biology: from genetic code to stochastic gene regulation. IET Syst Biol **4**(5), 311–329 (2010)
117. Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D.L., Falciani, F.: Modeling t-cell activation using gene expression profiling and state-space models. Bioinformatics **20**(9), 1361–1372 (2004)
118. Rau, A., Jaffrzic, F., Foulley, J.L., Doerge, R.: Reverse engineering gene regulatory networks using approximate bayesian computation. Statistics and Computing **22**(6), 1257–1271 (2012)
119. Ros, R., Hansen, N.: A simple modification in cma-es achieving linear time and space complexity. In: G. Rudolph, T. Jansen, S. Lucas, C. Poloni, N. Beume (eds.) Parallel Problem Solving from Nature PPSN X, *Lecture Notes in Computer Science*, vol. 5199, pp. 296–305. Springer Berlin Heidelberg (2008)
120. Sakamoto, E., Iba, H.: Evolutionary Inference of a Biological Network as Differential Equations by Genetic Programming. Genome Informatics pp. 276–277 (2001)
121. Samad, H.E., Khammash, M., Petzold, L., Gillespie, D.: Stochastic modelling of gene regulatory networks. Int. J. Robust Nonlinear Control **15**, 691–711 (2005). DOI 10.1002/rnc.1018
122. Savageau, M.A.: Biochemical systems analysis: Ii. the steady-state solutions for an n-pool system using a power-law approximation. Journal of Theoretical Biology **25**, 370–379 (1969)
123. Schäfer, J., Strimmer, K.: An empirical bayes approach to inferring large-scale gene association networks. Bioinformatics **21**(6), 754–764 (2005)
124. Schilstra, M.J., Nehaniv, C.L.: Bio-logic: Gene expression and the laws of combinatorial logic. Artificial Life **14** (2008)
125. Schmidt, M., Lipson, H.: Distilling free-form natural laws from experimental data. Science **324**(5923), 81–85 (2009)
126. Schramm, L., Jin, Y., Sendhoff, B.: Evolution and analysis of genetic networks for stable cellular growth and regeneration. Artificial Life **18**, 425–444 (2012)
127. Schwefel, H.P.: Numerical optimization of computer models. Chichester: Wiley & Sons (1981)
128. Seth, A.K.: A {MATLAB} toolbox for granger causal connectivity analysis. Journal of Neuroscience Methods **186**(2), 262 – 273 (2010)
129. Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulations network of escherichia coli. Nature Genetics **31**, 64–68 (2002)
130. Shin, A., Iba, H.: Construction of genetic network using evolutionary algorithm and combined fitness function. Genome Inform **14**, 2003 (2003)
131. Sîrbu, A., Ruskin, H., Crane, M.: Integrating heterogeneous gene expression data for gene regulatory network modelling. Theory in Biosciences **131**(2),

95–102 (2012)
132. Sirbu, A., Ruskin, H.J., Crane, M.: Comparison of evolutionary algorithms in gene regulatory network model inference. BMC Bioinformatics **11**, 59 (2010)
133. Sîrbu, A., Ruskin, H.J., Crane, M.: Cross-platform microarray data normalisation for regulatory network inference. PLoS ONE **5**(11), e13,822 (2010)
134. Sîrbu, A., Ruskin, H.J., Crane, M.: Stages of gene regulatory network inference: the evolutionary algorithm role. In: P.E. Kita (ed.) Evolutionary Algorithms. InTech (2011)
135. Solé, R.V., Valverde, S.: Are network motifs the spandrels of cellular complexity? Trends in Ecology and amp; Evolution **21**(8), 419 – 422 (2006)
136. de Sompel, H.V., Lagoze, C.: The Fourth Paradigm: Data-Intensive Scientific Discovery, 1st edn., chap. All Aboard: Toward a Machine-Friendly Scholarly Communication System, pp. 193–199. Microscoft Research (2009)
137. Spieth, C., Streichert, F., Speer, N., Zell, A.: Optimizing topology and parameters of gene regulatory network models from time-series experiments. In: K. Deb (ed.) Genetic and Evolutionary Computation GECCO 2004, *Lecture Notes in Computer Science*, vol. 3102, pp. 461–470. Springer Berlin Heidelberg (2004)
138. Spieth, C., Streichert, F., Supper, J., Speer, N., Zell, A.: Algorithms for modeling gene regulatory networks. In: Computational Intelligence in Bioinformatics and Computational Biology, 2005. CIBCB '05. Proceedings of the 2005 IEEE Symposium on, pp. 1–7 (2005)
139. Srinivas, N., Deb, K.: Multiobjective optimization using nondominated sorting in genetic algorithms. Evolutionary Computation **2**, 221–248 (1994)
140. Stricker, J., Cookson, S., Bennett, M.R., Mather, W.H., Tsimring, L.S., Hasty, J.: A fast, robust and tunable synthetic gene oscillator. Nature **456**, 516–519 (2008)
141. Swain, M., Hunniford, T., Mandel, J., Palfreyman, N., Dubitzky, W.: Modeling gene-regulatory networks using evolutionary algorithms and distributed computing. In: Cluster Computing and the Grid, 2005. CCGrid 2005. IEEE International Symposium on, vol. 1, pp. 512–519 Vol. 1 (2005)
142. Tang, K., Yao, X., Suganthan, P.N., MacNish, C., Chen, Y.P., Chen, C.M., Yang, Z.: Benchmark functions for the cec2008 special session and competition on large scale global optimization. Tech. rep., Nature Inspired Computation and Applications Laboratory (NICAL), China (2007)
143. Tegnèr, J., Yeung, M.K.S., Hasty, J., Collins, J.J.: Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. Proceedings of the National Academy of Sciences **100**(10), 5944–5949 (2003)
144. Thomas, S.A., Jin, Y.: Combining genetic oscillators and switches using evolutionary algorithms. In: Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on, pp. 28 –34 (2012)
145. Thomas, S.A., Jin, Y.: Evolving connectivity between genetic oscillators and switches using evolutionary algorithms. J. Bioinform. ComputBiol. **11**(3:1341001) (2013)
146. Thomas, S.A., Jin, Y.: Single and multi-objective in silico evolution of tunable genetic oscillators. In: R. Purshouse, P. Fleming, C. Fonseca, S. Greco, J. Shaw (eds.) Evolutionary Multi-Criterion Optimization, *Lecture Notes in Computer Science*, vol. 7811, pp. 696–709. Springer Berlin Heidelberg (2013)
147. Tobin, F.L., Damian-iordache, V., Greller, L.D.: Towards the reconstruction of gene regulatory networks (1999)
148. Tominaga, D., Okamoto, M., Maki, Y., Watanabe, S., Eguchi, Y.: Nonlinear numerical optimization technique based on a genetic algorithm for inverse problems: Towards the inference of genetic networks. In: German Conference on Bioinformatics'99, pp. 127–140 (1999)
149. Vignes, M., Vandel, J., Allouche, D., Ramadan-Alban, N., Cierco-Ayrolles, C., Schiex, T., Mangin, B., de Givry, S.: Gene regulatory network reconstruction using bayesian networks, the dantzig selector, the lasso and their meta-analysis. PLoS ONE **6**(12), e29,165 (2011)
150. Voit, E.O.: Model identifictaion: A key challenge is computational systems biology. In: Optimization and Systems Biology, pp. 1–12 (2008)
151. Voit, E.O., Almeida, J.: Decoupling dynamical systems for pathway identification from metabolic profiles. Bioinformatics **20**(11), 1670–1681 (2004)
152. Wang, Y., Zhang, X.S., Chen, L.: Optimization meets systems biology. BMC Systems Biology **4**(Suppl 2), 1–4 (2010)
153. Whitehead, D.J., Skusa, A., Kennedy, P.J.: Evaluating an Evolutionary Approach for Reconstructing Gene Regulatory Networks. In: Ninth International Conference on the Simulation and Synthesis of Living Systems (ALIFE9). MIT Press, Boston (2004)
154. Wieczorek, M.A., Jolliff, B.L., Khan, A., Pritchard, M.E., Weiss, B.P., Williams, J.G., Hood, L.L., Righter, K., Neal, C.R., Shearer, C.K., McCallum, I.S., Tompkins, S., Hawke, B.R., Peterson, C., Gillis, J.J., Bussey, B.: The constitution and structure of the lunar interior. Reviews in Mineralogy & Geochemistry **60**, 221–364 (2006)
155. Wiggins, C.: It's an exciting time for data in new york city. University of Columbia Engineering Newsletter (2012)
156. Xiao, M., Zhang, L., He, B., Xie, J., Zhang, W.: A parallel algorithm of constructing gene regulatory networks. In: D.Z. Du, X.S. Zhang (eds.) Optimization and Systems Biology, *Lecture Notes in Operations Research*, vol. 11, pp. 184–188. WORLD PUBLISHING CORPORATION (2009)
157. Xiong, J., Zhou, T.: Gene regulatory network inference from multifactorial perturbation data using both regression and correlation analyses. PLoS ONE

7(9), e43,819 (2012)
158. Yang, Z., Tang, K., Yao, X.: Multilevel cooperative coevolution for large scale optimization. In: Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on, pp. 1663–1670 (2008)
159. Yavari, F., Towhidkhah, F., Gharibzadeh, S.: Gene regulatory network modeling using bayesian networks and cross correlation. In: Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International, pp. 1–4 (2008)
160. Yip, K.Y., Alexander, R.P., Yan, K.K., Gerstein, M.: Improved reconstruction of ¡italic¿in silico¡/italic¿ gene regulatory networks by integrating knockout and perturbation data. PLoS ONE **5**(1), e8121 (2010)
161. Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D.: Advances to bayesian network inference for generating causal networks from observational biological data. Bioinformatics **20**(18), 3594–3603 (2004)
162. Yuan, Y., Stan, G.B., Warnick, S., Goncalves, J.M.: Robust dynamical network structure reconstruction. Automatica, Special Issue on Systems Biology **47**, 1230–1235 (2011)
163. Zhang, X., Zhao, X.M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J.K., Liu, Z.P., Chen, L.: Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. Bioinformatics **28**(1), 98–104 (2012)
164. Zhang, Z., Bajic, V.B., Yu, J., Cheung, K.H., Townsend, J.P.: Data integration in bioinformatics: Current efforts and challenges. In: D.M.A. Mahdavi (ed.) Trends and Methodologies, chap. 2, pp. 41–56. InTech (2011)
165. Zhao, S.Z., Liang, J.J., Suganthan, P., Tasgetiren, M.: Dynamic multi-swarm particle swarm optimizer with local search for large scale global optimization. In: Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on, pp. 3845–3852 (2008)
166. Zhu, H., Rao, R.S.P., Zeng, T., Chen, L.: Reconstructing dynamic gene regulatory networks from sample-based transcriptional data. Nucleic Acids Research **40**(21), 10,657–10,667 (2012)
167. Zitzler, E., Künzli, S.: Indicator-based selection in multiobjective search. In: in Proc. 8th International Conference on Parallel Problem Solving from Nature (PPSN VIII, pp. 832–842. Springer (2004)