Computational Approaches to the Modelling of Topological and Dynamical Aspects of Biochemical Networks

Adrián López García de Lomana

TESI DOCTORAL UPF / 2010

Supervised by: Jordi Villà i Freixa

Department of Experimental and Health Sciences

To my friends.

Acknowledgements

I need to thank many people who made this thesis possible.

First of all, I want to warmly acknowledge the expertise dedication, inspiration and enthusiasm of my PhD adviser, Jordi Villà i Freixa. He made possible that the bench biologist I was when I arrived to the group, become capable of typing sensible code: it would not have been possible without his patience. Sharing his scientific knowledge discussing ideas and motivating me has been a key for the success of the project. Without his constant support, this thesis would not exist.

During these years of my PhD, with three scientific stages abroad and collaborations with other groups, I had the opportunity to found exceptional colleagues that I want to mention here: Mike Hucka and Ben Bornstein at Caltech; Peter Coveney and Gianni de Fabritiis at University College London and Zoltán Oltvai and Qasim Beg at University of Pittsburgh. I specially should mention Berta Alsina for her patience and explanations.

I need also to mention many people I met in the lab: Montse, Karim, Michael, Àlex, David, Miguel, Pau, Marta, and Ignasi. It's been a pleasure to spend so many hours with you at work.

I don't want to forget the priceless help of the system administrators Alfons, Judith and Miguel.

Also I would like to acknowledge funds from AGAUR and HPC-Europa.

Finally, by simply being there, to my friends and family, thanks.

Abstract

Regulatory mechanisms of cells can be modelled to control and understand cellular biology. Different levels of abstraction are used to describe biological processes. In this work we have used graphs and differential equations to model cellular interactions qualitatively and quantitatively.

From different organisms, *E. coli* and *S. cerevisiae*, we have analysed data available for they complete interaction and activity networks. At the level of interaction, the protein-protein interaction network, the transcriptional regulatory networks and the metabolic network have been studied; for the activity, both gene and protein profiles of the whole organism have been examined. From the rich variety of graph measures, one of primer importance is the degree distribution. I have applied statistical analysis tools to such biological networks in order to characterise the degree distribution. In all cases the studied degree distributions have a heavy-tailed shape, but most of them present significant differences from a power-law model according to a statistical test. Moreover, none of the networks could be unequivocally assigned to any of the tested distribution.

On the other hand, in a more fine-grained view, I have used differential equations to model dynamics of biochemical systems. First, a software tool called `ByoDyn` has been created from scratch incorporating a fairly complete range of analysis methods. Both deterministic and stochastic simulations can be performed, models can be analysed by means of parameter estimation, sensitivity, identifiability analysis, and optimal experimental design. Moreover, a web interface has been created that provides with the possibility interact with the program in a graphical manner, independent of the user configuration, allowing the execution of the program at different computational environments. Finally, we have applied a protocol of optimal experimental design on a multicellular model of embryogenesis.

Resum

Els mecanismes de regulació de les cél·lules poden ser modelats per controlar i entendre la biologia cel·lular. Diferents nivells d'abstracció s'utilitzen per descriure els processos biològics. En aquest treball s'han utilitzat grafs i equacions diferencials per modelar les interaccions cel·lulars tant qualitativament com quantitativa.

En aquest treball s'han analitzat dades d'interacció i activitat de diferents organismes, *E. coli* i *S. cerevisiae*: xarxes d'interacció proteïna-proteïna, de regulació de la transcripció, i metabòliques, així com perfils d'expressió genòmica i proteòmica.

De la rica varietat de mesures de grafs, una variable important d'aquestes xarxes biològiques és la distribució de grau, i he aplicat eines d'anàlisi estadística per tal de caracteritzar-la. En tots els casos estudiats les distribucions de grau tenen una forma de cua pesada, però la majoria d'elles presenten diferències significatives respecte un model de llei de potència, d'acord amb proves estadístiques. D'altra banda, cap de les xarxes podrien ser assignades de forma inequvoca a cap distribució testejada.

Pel que fa a un nivell més microscòpic, hem utilitzat equacions diferencials per estudiar la dinàmica de models de diversos sistemes bioquímics. En primer lloc, una eina de programari anomenada `ByoDyn` ha estat creada des de zero. L'eina permet realitzar simulacions deterministes i estocàstiques, analitzar models mitjançant estimació de paràmetres, sensibilitat i anàlisi d'identificabilitat, així com dissenyar òptimament experiments. S'ha creat una interfície web que ofereix la possibilitat d'interactuar amb el programa d'una manera gràfica, independentment de la configuració de l'usuari, permetent l'execució del programa en diferents entorns computacionals. Finalment, hem aplicat un protocol de disseny experimental òptim en un model multicel·lular de l'embriogènesi en vertebrats.

Preface

This thesis should be understood within the personal background of the candidate. I graduated in Biochemistry in 2003. My daily work was in a *wet* lab, where my ultimate project consisted on the cloning and expression of a human protein in bacteria. Willing to approach biology from a more theoretical perspective, I started my PhD project in Barcelona in 2004. At that time, my deeper relationship with a computer was sending e-mails. As it can be understood, a new rich world popped up. Although at some points many hours had to be taken to understand the matter of difference of a single character in a long code, the possibilities and understanding provided by computer modelling has been largely rewarding.

The contributions to the field have been very specific at slightly separated areas of research. With the intention of putting into context the results of the thesis, a large broad introduction has been elaborated with the intention to bind together the three contributions. As a metaphor, the plot of the thesis is centripetal in which a global overview is necessary to arrive to the contributions, arranged in clearly precised areas. The large number of references raised as an aside consequence of such structure.

Contents

It always seems impossible until its done.

Nelson Mandela

1 Introduction

1.1 Background

A cell is a closed macroscopic entity of a structured arrange of molecules. With a typical size of 10 μm and 1 ng of mass, the complexity found within a single cell is stoning: just the number of proteins is estimated to be found in the order of 10^{10} molecules, of 10^4 different types [347], precisely located at different points of the subcellular structures. Additionally, proteins represent only about 20 % of the cell's weight [347]; it is hard to conceive the complexity resulting from the interactions with other macromolecules, like DNA for example. Complexity does not only come from the number of elements but also from their intricate interactions: dimerisation, poli-interaction, homo- and hetero-interactions, complex formation, activation, inhibition, gain of function, transport, and many more types of relations. And to make matters more inscrutable: space. Because the cell is rather different from an homogeneous soup or a simple compartmented chamber of nucleus and cytoplasm [370]. More than 20 types of different compartments are known in eukaryotic cells [9], each of them with different structure and function, clustering together molecules at specific environments. The structure of the cell provides, therefore, another independent level of complexity. Then, the quiet vision of homeostasis in the cell from an eye (and microscope) perspective is quite partial as could be the Earth from the outer space looking quiet and smooth while humankind's most complex behaviours appear from the large number of persons, physical and political barriers, cultures and

particular philias and phobias of each individual. Concomitantly to this view of a cell as a complex and dynamical system, it is understood that the description and knowledge of the cell molecular networks entitles a major step to comprehend the collective behaviour of macromolecules resulting on the cellular physiology.

1.2 Cellular Networks and Graphs

Down at the cellular level, several types of networks (Figure 1.1 **a.**) are subject of study. A typical example of a cellular network is the protein-protein interaction network (PIN) [464] , where proteins are the nodes of the network and links between nodes are defined as some kind of interaction (Figure 1.1 **b.**): ideally links depict protein physiological interactions but other relationships are typically described as results from yeast two hybrid (Y2H) experiments. Also, metabolic networks (MN) [169] can be constructed by linking two metabolites if they participate in the same reaction (Figure 1.1 **c.**). Other examples of cellular networks are the transcriptional regulatory network (TRN) [165] (Figure 1.1 **d.**) in which interactions consist on regulation of the expression of a gene by a transcription factor (TF). Other types of networks of biological interest are metabolic flux networks [10] or activity networks [568].

1.2.1 Graph Properties of Interest for Biological Networks

Biological networks can be represented as graphs [286]. Therefore, analytical tools provided by graph theory can be applied to the reconstructed set of interactions in order to understand key features from the biological networks. In general, graphs are defined as an ordered pair $G = (\mathcal{V}, \mathcal{E})$ were $\mathcal{V}$ is a set of vertices (also called nodes) and $\mathcal{E}$ is a set of edges (or links) which relate the vertices. We will use the notation $|\mathcal{V}|$ and $|\mathcal{E}|$ for the number of vertices and edges respectively. In this document I will use the term *node* as synonymous for *vertex*, the same way as *link* for *edge*.

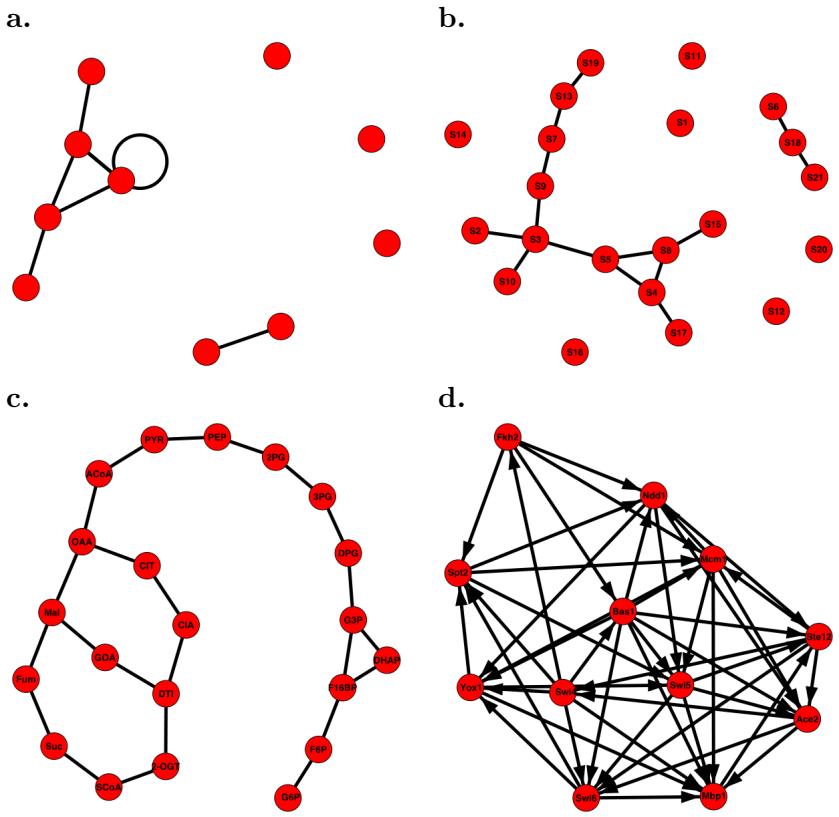

Figure 1.1: Types of networks. **a.** An example of the representation of a random network with 10 nodes and 9 links. Note the existence of self-interactions, sepated components and isolated nodes. **b.** An example of a PIN: nodes represent the proteins of the 30S subunit of the bacterial ribosome and links are the accounted physical contacts [532]. **c.** An example of a metabolic network: glycolysis and cytric acid cycle. Substrates are connected to each other through links that represent the actual metabolic reactions. Information retrieved from EcoCyc [295]. **d.** An example of a TRN: nodes represent the TFs responsible for the cell cycle in yeast [105].

Graphs can be defined of different type. When vertices are allowed to be linked only by a single edge, the graph is called *simple graph*; *multigraph* when multiple edges are present between nodes. When any vertex containing edges hold self-connections (loops) the graph is called a *pseudograph* (Figure 1.1 **a.**). Edges generally do not indicate direction and the graphs are termed *undirected graphs* (Figure 1.1 **b.**). If edges indicate direction, the graph is called *directed graph* (or *digraph*), being *oriented graphs* (Figure 1.1 **d.**) an special case of them, in which edges are restricted to indicate a unique direction (no bidirectional edges are allowed). Finally, other types of graphs commonly used to represent biological networks are *weighted graphs*, which associate a numerical value to each edge, generally quantifying the relationship between nodes, e.g., binding strength, frequency of the interaction or any other relevant attribute.

Several graph properties are relevant for the study of biological networks [227, 622]. Here I describe some of the local and global properties commonly used to describe biological networks:

- *Degree distribution:* Degree (k) of a vertex is the number of edges it has with other vertices of the graph. The degree of a vertex is considered a local property of the networks. The degree distribution $P(k)$ is the probability distribution of the degree of the vertices of a given graph. Commonly the degree distribution is presented as the cumulative degree distribution, this is, the probability of a vertex to have K edges being $K \geq k$. Degree distribution is one of the most popular measures to characterise a network [227]. In fact, several generative models of graphs can be distinguished by the degree distribution: graphs from either the Erdős-Rényi (ER) model or the Watts-Strogatz (WS) model [157, 592] display a Poisson degree distribution while the preferential attachment and growth model from the Barabási and Albert (BA) [43] renders a power-law distribution (see Section 1.2.2 for an explanation of these generative models). Important nodes of a network are those displaying large number of connections called hubs. In particular for biological networks, hubs tend to be essential for the survival of the organism [279, 611, 237, 612] although they do not show evolutionary constrains [48]. However, the biological interpretation of the as-

sociation between protein degree and essentiality is deeply rooted on the biochemical cellular processes rather than a simple correlation between degree and essentiality [627]. PIN hubs have been subjected to binomial classification as *party hubs* and *date hubs* [239]. The differences associated to this classification, which have been assigned to the node specific biological function rather than to the graph topology, seem to affect dramatically the topological properties. However this topic has caused recently large controversy [183, 182, 49, 50, 63] with latest works finding no evidence of support [3].

- *Clustering coefficient:* The clustering coefficient for undirected graphs is defined as a local property as

$$C_i = \frac{2e_k}{k_i(k_i - 1)}, \tag{1.1}$$

where C_i is the clustering coefficient of vertex i, vertex i is linked to k_i neighbours and e_k is the number of edges between the k_i neighbours. It provides an idea of the local cohesiveness of a vertex, the degree of tendency of forming groups or clusters. An average of C_i over all vertices $\mathcal{V}$,

$$C = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} C_i, \tag{1.2}$$

along with the shortest path length, gives us an idea of the *small-world* property[1] of a network. Another measure of interest is the $C(k)$ distribution which refers to the clustering coefficient of vertices of degree k. Different proposed generative models for biological networks present different $C(k)$ scaling. For example, hierarchical models (HM) [461] display an scaling law of $C(k) \sim k^{-1}$ while graphs resulting from the BA model present a constant value of $C(k)$ over the k range. However Soffer and Vázquez [526] explained that Equations 1.1 and 1.2 are biased towards degree correlations. Authors propose [526] another measure that takes into

[1] The *small-world* propertity is defined upon two concepts: small shortest path length and large clustering coefficient. The small-world network concept was firstly described by Watts and Strogatz [592]. They also reported [592] to find this behaviour in several natural, engineered and social networks, as the neural network of *Caenorhabditis elegans*, the power grid of western United Stated or a film actors network.

consideration the degrees of neighbour vertices:

$$\tilde{c}_i = \frac{t_i}{\omega_i}, \tag{1.3}$$

where ω_i is the maximum number of edges that can be drawn among k_i given the degree sequence of k_i, taking, consequently, the degrees of neighbour vertices into consideration. Using the new metric, clustering coefficient distributions for real networks became constant or logarithmic (as for other types of models [526]).

- *Distance:* A distance $d_{i,j}$ in a graph is defined as the minimum number of edges that connect vetex i with vertex j. The definition holds for non-weighted graphs, in the case of weighted graphs, the $d_{i,j}$ will be considered as the sum of the weights of that path. An interesting measure is the mean shortest distance (over all vertices $\mathcal{V}$ of the graph G),

$$\delta = \frac{1}{|\mathcal{V}|(|\mathcal{V}|-1)} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{V} \setminus \{i\}} d_{i,j}, \tag{1.4}$$

also called the mean geodesic distance [407][2]. This global measure is important for understanding the information flow in the network, centrality of nodes and network robustness against node loss. From the shortest path, other measures of the graph can be calculated like the eccentricity, which is the largest $d_{i,j}$ from a given vertex i to any other vertex of G,

$$\varepsilon(i) = \max_{j \in \mathcal{V} \setminus \{i\}} d_{i,j}. \tag{1.6}$$

[2] Newman defines the mean shortest path (following [407] notation) δ as:

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}, \tag{1.5}$$

where d_{ij} is the geodesic distance from vertex i to vertex j [407]. Newman includes for mathematical convenience the distances from each vertex to itself which are zero for definition. That makes the introduction of a minor error of factor $(n+1)/(n-1)$ negligible for practical purposes, that I avoid here. Another remark is that Newman only counts one half (and the diagonal) of the distance matrix for computational efficiency. Note the running index $i \geq j$. That is valid for undirected graphs as the set of shortest paths from i to j and j to i is the same. In Equation 1.4 we run over all pair of vertices but the diagonal, $i = j$.

The diameter of a graph is considered the largest eccentricity of G,

$$d_G = \max_{i \in \mathcal{V}} \varepsilon(i), \tag{1.7}$$

while the radius, the shortest eccentricity of G,

$$r_G = \min_{i \in \mathcal{V}} \varepsilon(i). \tag{1.8}$$

Additionally, Wuchty and Stadler [607] recapitulate other less popular measures like the status, the median or the Wiener index.

- *Centrality:* Centrality is a measure of the relative importance of a vertex or an edge to the network topology. Several variables can be used to determine centrality, like degree centrality, eigenvector centrality, closeness and betweenness [407]. The last two properties have been used to provide insights into biological networks [343, 341]. Given a vertex $i \in \mathcal{V}$, closeness, Cl_i, can be[3] defined as the reciprocal of the sum of the mean shortest distances between i and the rest of the vertices of the graph [486]:

$$Cl_i = \left[\sum_{j \in \mathcal{V} \setminus \{i\}} d_{i,j} \right]^{-1}. \tag{1.9}$$

Moreover, a very useful measure of centrality is betweenness [591],

$$B_i = \frac{1}{2} \sum_j \sum_k \frac{|D_{j,k}(i)|}{|D_{j,k}|}, \tag{1.10}$$

being $j \in \mathcal{V} \setminus \{i\}$ and $k \in \mathcal{V} \setminus \{i, j\}$[4]. $D_{j,k}$ refers to the set of shortest paths that go from j to k and $D_{j,k}(i)$ is the subset of $D_{j,k}$

[3] Other types of closeness have been described in [540, 419, 410, 118].

[4] In Wasserman *et al.* [591] betwenness is defined (following their notation) as:

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}. \tag{1.11}$$

Similarly to the case of ℓ [407], for computational efficiency only half of the matrix (in the current case without the diagonal, note $j < k$) is evaluated. As in Equation 1.4, for Equation 1.10 we do not consider only just one half of the distance matrix but the complete set of values but the diagonal ($i = j$) or the paths starting or ending at the node of reference, i.e., $i = j$ or $i = k$.

that contains vertex i. Defined for both, vertices [184] and edges [208], it is simply the number of shortest paths that pass through a given vertex or edge. A way to obtain a convenient measure of betweenness distributed on the range $[0, 1]$ is dividing B_i by the number of shortest paths evaluated:

$$B_i' = \frac{1}{(|\mathcal{V}| - 2)(|\mathcal{V}| - 1)} \sum_j \sum_k \frac{|D_{j,k}(i)|}{|D_{j,k}|}, \tag{1.12}$$

Edge betweenness can be used to detect community structure in a graph [208]. If a graph contains clusters of vertices highly connected while still connections are sparse between the clusters, the graph is said to display community structure. Social and biological networks are known to display community structure [208]. Interesting information about the structure of the network and the behaviour of the nodes can be extracted from the knowledge of community structure. Specifically for biological networks, two functional aspects of the network can be studied provided the community organisation: (1) nodes of a given community are more likely to share certain properties, as for example function and (2) specific links and nodes responsible for connecting the scattered communities are likely to be key players of the network as their loss or malfunction will affect dramatically the whole behaviour of the system. Specifically for cellular networks, those *connecting* nodes correlate with top regulators, conserved sequences or house-keeping genes [343, 341].

- *Community Structure:* As mentioned above, important information from the network can be derived from the study of the community structure of a graph in general [235, 442]. However, a large variety of graph properties can be studied to determine community structure in a network, rendering different results at different computational efforts[5]. Apart from various methods I highlight afterwards, many others have been described [472, 432, 166, 515, 352, 266, 265, 336, 363, 382] and I am confident that many others will be devised in the close future. I suggest the excellent review by S. Fortunato for a through exposition of the topic [179]. Here I introduce some of the most commonly used methods to study community structure with special emphasis in biological networks:

[5]Not necessarily the most computationally expensive methods provide the best results.

– *Betweenness:* As introduced before, Girvan and Newman [208] developed an algorithm to find community structure in graphs. The idea of the algorithm is the following:

  1. For each edge, calculate its betweenness value.
  2. Remove the edge of the graph with the highest value of betweenness.
  3. Go to step 1. until all edges have been removed.

  Displaying the results of the removals as a hierarchical tree will reveal the potential communities. The idea behind is that edges connecting communities have high values of betweenness and removing them will reveal the underlying communities. Two essential elements of the algorithm need to be highlighted for adequate results: (1) edge betweenness [208] and not vertex betweenness [184, 216, 285] should be calculated and (2) recalculation of the betweenness values should be done after the removal of an edge. Not all edges of the graph necessarily variate their betweenness when a single edge is removed, but if the values of betweenness of the original graph are used for the trimming, the community structure will not be revealed. The original algorithm, and modifications of it, have been applied to biochemical networks (metabolic and whole-cellular networks [148, 261, 90] and gene interaction networks retrieved from bibliographic sources [598]), finding community structure at different levels of the graph hierarchy, as well as a general trend in social networks [19, 414, 230]. Additionally, very similar measures to betweenness like information centrality [181] have also been used within this framework with similar results. Finally, vertex betweenness has been shown to correlate with essentialilty and evolutionary age in PINs [285].

– *Modularity:* Defined by Newman as [408, 414, 409, 231]

$$Q = \sum_{i \in \mathcal{Z}} \left(\frac{l_i}{|\mathcal{E}|} - \left(\frac{d_i}{2|\mathcal{E}|} \right)^2 \right) \tag{1.13}$$

  can be used to help algorithms to find the optimal partitioning of a graph into communities [102]. Given a graph partitioned into several communities $i \in \mathcal{Z}$, a high modularity value

should be observed if the number of edges within the community is larger than the *expected*. By *expected*, we consider the number of edges of an ensemble of vertices connecting with equal probability between them and to the rest of the vertices of the graph. The ensemble would be of the same size as community i, showing the same degree sequence. Therefore l_i is the number of edges that link vertices of community i and d_i is the sum of degrees of vertices of community i.

This measure Q has been used as the minimisation function for graph partitioning using deterministic [327] or heuristic methods, like simulating annealing or Monte Carlo sampling optimisation [533]. Guimerà and Amaral [231] have applied that protocol to metabolic networks discovering relevant biological facts, like that non-hub metabolites connecting different modules are more conserved than hubs that connect metabolites of the same module. Other optimisation algorithms, like extremal optimisation [71], have been applied with the same purposes, the maximisation of Q [146]. Furthermore, other authors combined optimisation and spectral analysis [450] to obtain graph partitions with larger Q than other methods [208, 409, 146]. An interesting review comparing sensitivity and performance for several methods, mainly based on modularity, has been published by Danon *et al.* [119]. Worth noting is the fact that random networks [157] and networks with power-law degree distributions [43] display relatively large values of modularity [232]. In order to assign a *significant* value of modularity to the network in hands, Guimerà *et al.* [232] propose the comparison with the corresponding null model.

However, modularity, Q, poses some problems when used as a minimisation function for community structure detection. Fortunato and Barthélemy [180] proved that the maximisation of Q might provide with a specific partition of the network forced by the algorithm, not reflecting the actual structure of the network. Authors proved that the number of communities $|\mathcal{Z}|$ that maximises the modularity Q of a graph $G(\mathcal{V},\mathcal{E})$ is $|\mathcal{Z}|^* = \sqrt{|\mathcal{E}|}$, derived from the equation,

$$\frac{\mathrm{d}Q(|\mathcal{Z}|,|\mathcal{E}|)}{\mathrm{d}|\mathcal{Z}|} = -\frac{1}{|\mathcal{E}|} + \frac{1}{|\mathcal{Z}|^2}. \qquad (1.14)$$

At this point it is clear that communities derived from the maximisation of Q are forced to have a certain size. Those communities, specially the loose ones ($l_i \approx \sqrt{2|\mathcal{E}|}$ or lower), are likely to be detected upon the aggregation of *actual* communities. Consequently, authors re-analysed some of the most studied networks so far, some biological, like the TRN of *S. cerevisiae* or *E. coli* and the neural network of *C. elegans* and some others like social or electric networks. They found that most of the communities previously described [231] exhibited themselves community structure. Therefore, a word of caution should be given when the community structure has been inferred from the optimisation of the modularity, displaying, likely, a poor resolution of the communities.

Resolution limit problems have been solved by an elegant and simple solution proposed by Arenas *et al.* [21]. The idea consists on the introduction of a self-loop of variable strength r to each vertex of the graph. Using optimisation methods (extremal optimisation [146] or tabu search) at different values of r, the community structure resolves at different levels of granularity. Authors explore a wide interval of r, including the complete meaningful range, observing at different levels of r, different graph partitions, some of them showing larger stability than others.

A complementary alternative is the calculation of the local modularity [396], where only vertex neighbours are taken into consideration to calculate the modularity. This way the resolution of the communities is much larger and in particular for biological networks, biological function is inferred more confidentially. In this direction, Medus and Dorso [374] take the weak and strong community definitions [453] as a *merit factor*, i.e., as a function subject of optimisation. In particular they use a simulated annealing protocol. Consequently, using a local approach for modularity proxy, the community resolution problem is surmounted. Moreover authors use specific graph benchmarks [319] to provide proofs for a better performance than Equation 1.13.

Other works [483] have also efficiently solved the problem of community resolution limit using other different approaches.

– *Hierarchical Structure:* Hierarchical random graphs have been proposed as a network model [101]. They are based on the connection of two vertices depending on the degree of relatedness. Using maximum likelihood estimation and Monte Carlo sampling, hierarchical random graphs have been matched with empirical biological networks, like the metabolic network of *Treponema pallidum*. Using these models which inherently incorporate communities, several critical features have been determined as missing links or false-positive node associations [101].

  Alternatively, Sales-Pardo *et al.* [493] have developed a different method that retrieves very accurately nested hierarchical structure of real complex networks as those representing flight connections, metabolism or e-mail communications. The accuracy they show retrieving world-wide socio-political boundaries using just information about flight connections is impressive. The method they propose proceeds as follows: first, node affinities based on the complete modularity landscape are calculated to assess node grouping. Then, the specified groups are compared with a random network to assess whether or not the defined modules contain themselves modularity structure. Repeating iteratively this last step, modularity at the different levels of the hierarchy is uncovered. Furthermore, using simulated annealing and a cost function, a rearrangement of the groups of nodes is performed so that groups of larger affinity are closer. Finally an optimal number of modules at each hierarchical level is determined using a Bayesian information criterion.

– *Loops:* Radicchi *et al.* [453] devised an algorithm based on the *edge-clustering coefficient*, which is a local measure for each edge $(i,j) \in \mathcal{E}$ so,

$$C_{(i,j)}^{|L|} = \frac{z_{(i,j)}^{|L|} + 1}{s_{(i,j)}^{|L|}}, \tag{1.15}$$

  where (i,j) is the edge that links vertices i and j, $|L|$ is the size of the loop for which we are calculating the measure (loops of size 3 and 4 have been studied), $z_{(i,j)}^{|L|}$ is the actual number of

loops of size $|L|$ that include (i, j) and $s_{(i,j)}^{|L|}$ is the maximum possible number of loops of size $|L|$ to which (i, j) belongs. Basically, it consists on the ratio of cycles of size $|L|$ based on each edge with respect to the maximum possible. The measure anti-correlates with betweenness for some empirical networks studied by the authors. Radicchi *et al.* follow the protocol of Girvan and Newman [208], but using the edge-clustering coefficient instead of betweenness. The results are very similar but slightly more accurate than for the Girvan and Newman algorithm [208]. Moreover, the computational cost is cheaper, scaling between $\mathcal{O}(|\mathcal{V}|^2)$ and $\mathcal{O}(|\mathcal{V}|^3)$ while the Girvan and Newman algorithm [208] scales $\mathcal{O}(|\mathcal{V}|^3)$. Finally, they propose a useful measure to stop the algorithm based on the comparison of the structure of the current subgraph with a random graph of vertices with the same degree sequence. They use that metric to stop the algorithm and retrieve naturally the size of the communities. However, one of the major drawbacks of the algorithm is that the closed loops are frequently found in assortative networks (social networks) and authors claim that the application of the algorithm at disassortative networks (technological or biological networks) can be problematic.

– *Cliqueness:* In principle, vertices of a graph can belong to different communities simultaneously. Within the biological context, in fact, pleiotropic proteins are typically involved in several pathways or can display different functions in different cell types. However, community structure algorithms have generally eluded this issue and most of the algorithms classify vertices unequivocally to a single community. Palla *et al.* [424, 131, 1, 425] bring this problem into consideration and they study three complex networks, the *Saccharomyces cerevisiae* PIN among them. While retrieving an accurately distribution of proteins into different communities corresponding to their biological function, surprisingly they found that 26 % of the proteins belong to more than one community, proving the importance of nodes shared by different communities. Their algorithm is based on cliques, which are subgraphs where each vertex is connected to all other vertices of the subgraph. The

authors used k-cliques and based their results on k-clique adjacency, meaning that two cliques are adjacent if they share $k-1$ vertices. Then, a k-clique community [425] is defined as the sub-graph resulting from the union of adjacent k-cliques.

– *Spectral analysis:* Spectral properties of the connectivity matrix of a graph can be exploited to uncover the community organisation [74, 78, 140, 81, 515]. Newman [413, 412] has developed an spectral method improving previous results [208, 102, 146] (in accuracy, maintaining performance) based on the spectral properties of the network modularity matrix, $B_{i,j}$, defined as

$$B_{i,j} = \sum_{\{i,j\}\in\mathcal{V}} A_{i,j} - \frac{k_i k_j}{2|\mathcal{V}|}, \tag{1.16}$$

being $A_{i,j}$ the adjacency matrix, k_i the degree of vertex i and $|\mathcal{V}|$ the total number of edges. Concomitant refinement of the results, inspired on the Kernighan-Lin graph partitioning algorithm [294], also proposed by Newman, provide a final communities vertex re-distribution that optimises the modularity of the graph divisions. Other algorithms very similar to the Newman proposition based on the Kernighan-Lin algorithm have been tested. Mei *et al.* [375] use an algorithm based on single-node-move operations starting from a random partition as initial condition that outperforms for the Q value the complete spectral analysis protocol [413] and other methods [483, 327].

A nice property of the Newman algorithm [413] is that the community partitioning is stopped automatically and it provides with a procedural definition of *community* as the indivisible subgraph [413]. Alternative efforts based also on spectral properties have been proved to be successful to retrieve community structures in graphs. For example Richardson *et al.* have extended the analysis to spectral tripartitioning [471]. Furthermore, Arenas *et al.* [20] have related spectral properties of the Laplacian matrix to the dynamics of group oscillators. Authors took advantage of synchronisation in coupled oscillators to detect correlated dynamics of vertices and classify them into communities.

- *Statistical mechanics:* Reichardt and Bornholdt [465, 466] introduced a formal framework derived from statistical mechanics to find community structure in networks. Their approach connects ideas from the Potts model with the well-known modularity measure in Equation 1.13. Concepts as vertex cohesion and adhesion to a given community are introduced. The algorithm is able to detect overlapping communities, furthermore, tuning a single parameter their algorithm is capable to find embedded hierarchical structures. Finally, an interesting feature of the method is the capacity to determine to which community a given vertex belongs to without classifying the whole graph, a useful feature for large networks. Recently their algorithm has been applied to PINs [330].
- *Statistical methods:* Maximum likelihood estimation has been used to assign vertex association to a given group. As for the work of Reichardt and Bornholdt [465, 466], Hastings [242] converted the graph partitioning problem into a maximum likelihood inference problem, linking the Potts model and belief propagation. However one of the major drawbacks of the method is that it takes the number of communities as an input. Furthermore, not many types of networks have been tested against the proposed method.

  Independently, Newman and Leicht [415] developed a very effective method to uncover complicated relationships of associativity and it works perfectly for both assortative, disassortative and neutral networks. Moreover, a probability is given for each vertex to belong to a community which provide a valuable information about how strongly that vertex belongs to that community. This feature is very convenient because communities are not always sharply separated and certain vertices can shown inclination towards several communities. Again a drawback of the method is that the number of communities has to be provided by the researcher. Several partitions can be tested for a maximum value of likelihood but the procedure could be cumbersome for large networks.
- *Processes propagation:* Wu and Huberman [601] take the original idea of viewing a graph as a an electric circuit. First, each vertex is understood as a resistance. Then a voltage is applied

to any two vertices of the graph and thanks to the Kirchhoff equations, the voltage to each node can be calculated. Analysing the spectrum of voltages, vertices can be classified within different communities. One the main advantages of the method is that it runs in $\mathcal{O}(|\mathcal{V}|)$ time, which makes it one of the fastest methods for graph community finding. On the other hand, some drawbacks are present. For example, vertices at which voltage would be applied (the *poles*) should belong to different communities for a correct functioning of the algorithm. Authors provide useful stratagems to solve this issue. However, another more serious problem is, again, that the number of communities should be specified *a priori*.

On the other hand, Nandini *et al.* [454] propose an algorithm based on label propagation that overcomes those problems. The simple algorithm, which also runs in $\mathcal{O}(|\mathcal{V}|)$ time, starts with the labelling of each vertex with a unique label. Each vertex is then updated asynchronously in such a way that it takes the most frequent label among its neighbours. Labels propagate along the graph, stopping the algorithm when each of the vertices of the graph is labelled with the most frequent label among its neighbours. This simple and fast algorithm, provides with very good partitions, comparable to other algorithms much more computationally intensive. Moreover, no assumption is taken about the number of communities, it is the underlying structure of the graph itself which guides to it.

– *Entropy-based measures:* Several authors [481, 64], have provided with different alternatives to Equation 1.13 to assess significance of community structure in networks. Based on entropy, authors have applied their methods to social, spatial and biological networks showing in some cases a better behaviour than Equation 1.13.

Once vertices are assigned to communities in a graph and a community value has been assigned, we should consider if that value could be explained by chance or not: is the graph truly modular? A first and intuitive approach would be to calculate a z-score from an ensemble of null model graphs. However Karrer *et al.* [291] have provided with a much more elegant and reliable method based on information theory measurements. The protocol consists on a fair

perturbation in the arrangement of edges and the measurement of the variation of information of the two classifications based on the conditional Shannon entropy. Moreover, authors prove that the use of the z-score may lead to the misinterpretation of the results given that very high z-score values are not always consistent with strongly modular graphs [291].

- *k-core structure:* Similar to community structures, graphs vertices can be classified into k-core structures. k-cores were introduced by Seidmann [513]: given a graph $G = (\mathcal{V}, \mathcal{E})$, a k-core is defined as the maximum subgraph $G_k = (\mathcal{V}_k, \mathcal{E}_k)$ such that each vertex $v \in \mathcal{V}_k$ has at least degree k. k-core classification is a valuable tool for coarse grain analysis of graphs, it might not reveal the fine structure of small dense associations but it proved to be a very helpful tool for the visualisation of large networks [11]. k-core decomposition has been applied to several biological networks resulting on interesting insights about protein evolution and essentiality [605, 606].

- *Connectivity indexes:* Based on distance, several connectivity indexes can be calculated:

  - *Matching index:* The matching index, $M_{i,j}$, is of special importance on the study of biological networks as it can provide with a method to reveal nodes of the network that share function. In PINs, two proteins that participate on the same function does not necessarily bind together but certainly they will share common proteins to which they connect. The matching index is defined as the number of shared neighbours of i and j divided by the total number of neighbours of i and j [286]. Nodes with a high matching index are more probable of being involved on the same biological function.

  Other interesting connectivity indexes, less common in the study of biological networks are:

  - *β-index:*

$$\beta = \frac{|\mathcal{E}|}{|\mathcal{V}|}, \tag{1.17}$$

    which provides an idea of the connectivity of a graph comparing the number of edges and number of vertices.

– γ *index:*

$$\gamma = \frac{|\mathcal{E}|}{\frac{1}{2}|\mathcal{V}|(|\mathcal{V}|-1)}; \tag{1.18}$$

which compares the actual number of edges against the maximum number of edges possible.

- *Edge mixing pattern:* Alternative variables can be associated to the nodes of a network. These variables can be either inherent to the graph, like the connectivity degree distribution $P(k)$, or others, alien to the mathematical representation. For biological networks, for example, a biological function can be associated to each node; in the case of social networks, the income, race or speaking language; etc. Using those variables as classifiers, it can be observed whether if elements of the same group tend to bind together (assortative mixing) or to avoid each other (disassortative mixing). A commonly used index is the assortativity coefficient [406],

$$r = \frac{\mathrm{Tr}\ \mathbf{e} - ||\mathbf{e}^2||}{1 - ||\mathbf{e}^2||} \tag{1.19}$$

where $\mathbf{e}$ is the matrix whose elements $e_{i,j}$ represent the fraction of edges that connect vertices of type i with vertices of type j. Tr accounts for the trace of a matrix and $||\mathbf{x}||$ is the sum of the elements of $\mathbf{x}$. Generally speaking social networks are assortative while technical and biological networks are disassortative [405, 365, 406]. Interestingly, some important properties of graphs vary depending on the $P(k)$ associative mixing coefficient [405, 406]. Positive assortative mixing coefficient graphs display more robustness than neutral or disassortative graphs against hub removal which is an important property in networks of different fields: it contributes to the control of epidemic outbreaks, it determines system failures in targeted attacks of computer networks or it predicts gene knockout resilience in biological systems.

- *Subgraph Frequencies:* An interesting local measure is the frequency at which certain small subgraphs appear in the whole graph. Those subgraphs that appear more frequently than expected are called motifs [386]. Several motifs have been described in PINs [386, 385] relating the topology to the biological role through dynamical analysis [517, 480, 479, 359, 361, 289, 360]. Further studies [447] have

concluded that the general distribution of frequencies of motifs is closer to geometric random graphs [431] than other types of generative models like Erdős-Rényi [157, 158, 159] or small-world networks [592] (see Section 1.2.2 for an explanation of the generative models). The relative distance between two graphs, G and H, can be calculated [447] based on the difference of frequency of a set of subgraphs, $g \in \mathcal{S}$, as

$$D(G,H) = \sum_{g \in \mathcal{S}} \text{abs}\Big(\nu_g(G) - \nu_g(H)\Big), \tag{1.20}$$

being $\nu_g(G) = -\log\left(\frac{|g|}{\sum_{g \in \mathcal{S}} |g|}\right)$, accounting for the relative frequency of subgraph g in G.

Other graph measures have been used to describe biological and other types of networks. Some of them are network efficiency [321], vertex accessibility, spectra properties derived from the adjacency matrix [65, 164], local connectivity measures [234, 384], rich-club phenomenon [106] and many others. Finally, I suggest the reading of [407], an excellent review about the study of network properties. We are confident that new graph measures will be described in the following years providing us with further description and deeper understanding of biological networks.

1.2.2 Graph Generative Models for Biological Networks

As some of the main graph properties employed to study biological networks are defined, now a new question arises. Can we model the whole network? Are we able to understand how the network was constructed, under which principles? On the way to answer these fundamental questions, several generative models have been proposed, both for networks in general and specifically for biological networks too.

Generative models represent a key tool for the understanding of the tenets of the studied networks. For example, in the biological field, some models

provide good approximations for the degree distributions of the empirical networks, other models work well to simulate the observed clustering coefficient, some others reproduce faithfully intricate properties as subgraph frequencies, and so on. However, unfortunately, no single model has fulfilled yet all properties displayed by graphs representing biological networks.

Some of the most popular models used are the following:

- *Random graphs:* Random graphs were formalised and analysed by Erdős and Rényi [157, 158]. A random graph $G(|\mathcal{V}|, p)$ [203] is a random graph if it belongs to the ensemble $\mathcal{G}$ such that the $|\mathcal{V}|$ number of vertices are connected with independent probability p. Therefore, the probability for a random graph to have precisely $|\mathcal{E}|$ edges is $p^{|\mathcal{E}|}(1-p)^{M-|\mathcal{E}|}$ where M is the maximum number of edges possible, $M = \frac{1}{2}|\mathcal{V}|(|\mathcal{V}|-1)$. In other words, the expected number of edges is $\frac{1}{2}p|\mathcal{V}|(|\mathcal{V}|-1)$ and the expected mean degree is $\langle k \rangle = p(|\mathcal{V}|-1)$. Random graphs can alternatively be defined specifying directly the number of vertices and edges, as $G(|\mathcal{V}|, |E|)$.

  Many properties have been studied for random graphs [72, 277, 290]. The most relevant one is the phase transition from low to high p values. Defined a phase transition point in $z = \langle k \rangle = p(|\mathcal{V}|-1)$, if $z < 1$, connected components are typically small, with values no larger than $\mathcal{O}(\log |\mathcal{V}|)$ while for $z > 1$ a giant component appears holding $\mathcal{O}(|\mathcal{V}|)$ of the vertices. At both cases distribution of components' sizes is exponential. At the phase transition point, $z = 1$, however, the largest component is of size $\mathcal{O}(|\mathcal{V}|^{2/3})$, with a power-law distribution of components' sizes characterised by a scaling exponent of 3/2. Other properties typical of random graphs from the Erdős and Rényi (ER) model are Poisson distributions for the degree distribution, low clustering coefficient ($C \approx \bar{k}/|\mathcal{V}|^{-1}$) and low mean shortest distance ($\delta \approx \log |\mathcal{V}|/\log \bar{k}$) [72].

  Although most real networks do not behave as random graphs, the latter represent fairly a first approximation of the former. Moreover, they can be used to assign specific features to real networks. Used as the null model [466], real network properties that deviate from the ones of random graphs can be said to be significantly relevant at the system studied. On the case of biological networks, de-

gree distributions are far from Poisson distributions, although some properties like the mean shortest path is closely approximated by random graphs.

One step further is provided by random graphs with specific degree distributions [58], power-law, for example, which have been proved to provide a very exact approximation to some real networks [416]. Finally, hierarchical random graphs [101] have been defined recently.

- *Small-world graphs:* Watts and Strogatz [592] defined a type of graphs called *small-world*. The idea is the following. First start from a regular lattice (the most studied case is for lattices of one dimension although higher dimensions have been used too [395, 403, 417, 127, 423]) where a vertex is linked to k neighbours with periodic boundary conditions. Then with probability p re-wire each edge to another random vertex so no self-edges nor double edges appear. Moving from a regular lattice ($p = 0$) to a random graph ($p = 1$), there exists a region (Figure 1.2) where the resulting graph is a small-world graph. Apparently, the graph is much more similar to a regular lattice than to a random graph, due to the low number of new edges, but the properties of the resulting graph are drastically different [45, 404, 542]. Basically, those new edges are "shortcuts", that increase dramatically the connectivity of the network and while the clustering coefficient remains very similar to the original regular lattice value, the mean shortest path decreases to the values of random graphs. Recently quantitative measures that evaluate the degree of *small-world-ness* of a graph have been presented [269]. Small-world models have been very successful to explain many real systems, including biochemical networks [592, 13, 172, 280, 587, 527, 353]. However, particularly at metabolic networks, some concerns have been exposed about the network representation [391] given that analyses of more biochemically sensitive representations fail to disclose small-world properties [22].

- *Growth models:* The two previous models presented here did not consider how networks may evolve. On the other hand, several other network models take into consideration the starting point of few nodes and links and establish some rules for the addition of

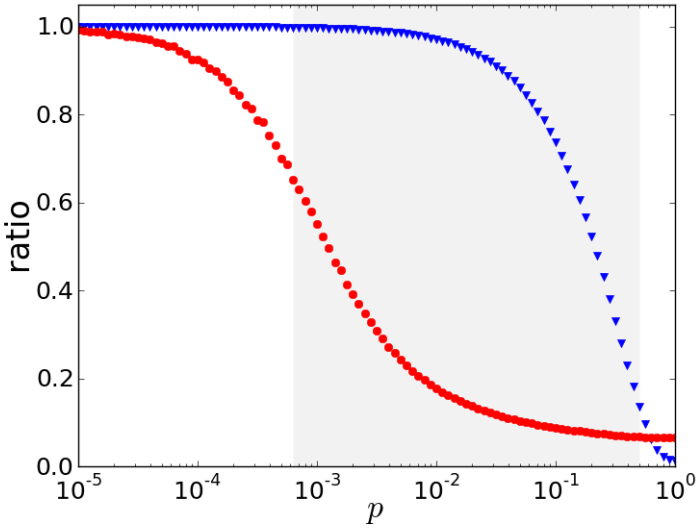

Figure 1.2: Starting from a regular lattice (ring topology) of 1,000 vertices, each connected to $k = 10$ other neighbouring vertices, with probability p, each edge is rewired randomly. Therefore at $p = 0$ we have a regular lattice while at $p = 1$ the graph is a random network. Typically for a regular lattice, clustering coefficient and mean shortest path are large while for a random graph, both measures are low. In this figure I depict the relative values with respect to a regular lattice of the clustering coefficient (blue triangles) and the mean shortest path (red dots). Each value has been averaged over 1,000 graph instances. Interestingly, along the p-axis there is a broad range (grey frame) at which the clustering coefficient remains within the same order of magnitude of that of a regular lattice, while, at the same time, the mean shortest path is already within the same order of magnitude of a random graph: the graph is a small-world.

new elements.

In the early 1960s, Price [444, 445] applied ideas of Simon [520] to describe and explain networked systems. Simon showed that power-law distributions are displayed by economic systems where the rule "the rich gets richer" is applied. Price put this knowledge in the context of bibliographic networks and he provided with the first explanatory mechanism for power-law degree distribution of networks of real systems. The mechanism consists on the addition of a newly introduced vertex of defined m degree to the network such that each edge of the new vertex is added to the old vertices with probability proportional to the degree of the old vertex.

Later Barabási and Albert starting from the same ideas, applied them to a growth model for the World Wide Web [43] and coined the new term, *preferential attachment*[6]. Their model consists on two simple rules: growth and preferential attachment. Starting from a bunch of initial vertexes m_0 (typically a low number like 1, 3, 5 or 7), a new vertex is introduced, with m edges ($m \leq m_0$). Each of the m edges are bound to vertex $i \in \mathcal{V}$ with a probability proportional to its degree as following:

$$p(i) = \frac{k_i}{\sum_{j \in \mathcal{V}} k_j} \tag{1.21}$$

An important difference from the Price model is that the Barabási-Albert (BA) model is defined on undirected graphs (while Price worked with directed networks) which provided an excellent work-around to treat formally the growth process. Much attention within the scientific community has been drawn to the BA model and, in fact, graphs created using the BA mechanism display power-law degree distributions, which have been claimed to be observed in many real systems [2, 463, 411, 121], including biological networks [280, 279, 451, 527, 571, 152, 10, 568, 42, 6]. The BA model can be easily understood from an evolutionary biology point of view as a process of gene duplication [299, 527, 100, 427, 571, 585, 61]. Additionally, the inclusion of other mechanisms to the BA model (different forms of biological divergence) can render graphs holding other

[6] Price called it *cumulative advantage* and it is also known as the *Matthew effect* [380] in sociology.

types of scaling on the degree distribution [7, 527, 570, 571, 427] like truncated power-laws, logarithmic or multifractal. As a matter of fact, many biological constrains have been postulated to account for the truncated power-law degree distributions like molecular crowding, aging or molecular surface drain [142, 13, 307]. Nevertheless, graphs with the same degree distribution may display very different properties [143] concerning clustering coefficient, degree correlation, performance, robustness, etc. For example, a graph of 100 triangles has the same degree distribution that a single 300-nodes cycle, but completely different local properties [449]. Additionally, not all real networks show power-law forms for their degree distribution [13] and some features from real networks are not adequately modelled by the BA mechanism [407, 554, 285, 383, 302]. Moreover, the BA model is not the only one that yields power-law degree distributions; some references, among many others are: [416, 79, 141, 4, 143, 285, 411, 492, 62, 129, 387, 67, 195]. Still, the BA model is a popular generative network model for biological networks.

Another growth mechanism of interest for biological networks is the hierarchical model [461, 460]. Starting from an small cluster of n connected vertices with a central vertex, the cluster is copied $n-1$ times and the peripheral vertices of the new replicas are connected to the central vertex of the original cluster. Repeating this process several steps, a hierarchical graph is created with power-law degree distribution and $C(k) \sim k^{-1}$ scaling [7] for the clustering coefficient. This model provides a better approximation than the BA model for the clustering coefficient of metabolic networks. Other aggregation generative models provide too similar results on the degree distribution and the clustering coefficient [141, 101].

- *Geometric random graphs:* Recently, geometric random graphs (GRGs) [431] have been provided as a striking good approximation for local measures (subgraph frequencies [447] and graphlet degree distribution [446]) in biological networks, specifically in PINs [447, 446, 443, 252]. GRGs can be constructed by placing randomly points in

[7] Soffer *et al.* [526] have shown that the decreasing scaling is a consequence of degree-correlation biases. Without the bias, empirical networks show constant or logarithmic scales (constant for PINs). Authors provide with a new clustering coefficient index, free of vertex degree correlations, see Section 1.2.1.

a given metric space. Those points closer to a threshold are bound. A variety of graphs can be obtained using different N dimensional spaces, different types of distances (Euclidean, Hamming, etc.), and thresholds. It results very interesting that although the definition of GRGs is so simple, trained-GRGs [316] are very capable to accurately describe detailed features such as subgraph frequencies, additionally to the degree distribution, the clustering coefficient or the graph diameter. Moreover, from an evolutionary point of view, GRGs can be constructed using simple mechanisms. The resulting graphs provide with a better similarity than other classical protein evolution models on structural local aspects [449]. Finally, GRGs have been used to assess false positives and false negatives in PINs [317].

Other generative models have been shown to be useful to study specific aspects of biological networks [570, 448, 387]. Finally, other generative models from other fields like firm growth [536, 326] or airport networks [195] could be adapted to biochemical networks.

1.2.3 Degree Distribution Determination

As it has been shown previously, different generative models may render different degree distributions. A first and obvious step on the validation of network models is to determine if generative models can explain, among other features, the observed degree distributions. Several methods have been used so far, some of them presenting problems or incompatibilities.

Firstly, merely from the data acquisition, partial sampling of actual networks, due to the lack of complete coverage from experimental limitations, has raised serious problems [545, 237] about the conclusions that can be inferred. More generally, it has been suggested [296, 543, 550, 602] that the charactheristic observation that biological networks follow a power-law distribution may have been reached due to methodological shortcomings. Specifically, it has been argued that analysing relatively small cellular networks, having only a few hundred to a few thousands data elements using frequency-degree or intensity plots, does not have

sufficient power to differentiate among various network models having heavy-tailed distributions, and that the use of rank-degree plots proves superior for this purpose [296, 543, 550, 602].

As explained in Section 3.2.3, the assessment of the best model explaining a given data distribution has typically been done using simple linear regression methods. These methods are suitable for normal distribution functions, but not for highly skewed distribution functions. Essentially, the problem arises from the fact that skewed distributions are characterized by the scale of the tail, which forms most of the support for the distribution but it is barely populated (i.e. contains less than 10% of the data points). Because of this, simple least square fits of the probability density distribution computed via histogram methods are very poor estimators of the distribution parameters (see [411, 550] for a review of possible problems) due to the noisy poor sampling of the tails.

Therefore, it has been argued that density plots should not be used as a base for the fitting of these types of data, as several more reliable methods are available. A simple and better strategy is to use rank-plots as commonly used in engineering and economics [550]. Logarithmic binning [8] has also been used as a more robust alternative, but it has been reported that this procedure fails to retrieve the value of the exponent as the slope of the graph for power-law distributions [550, 219, 103]. Finally, a logarithmic transformation could be applied to the data and fit the corresponding distribution function [335] thus avoiding the problem of skewed data from the outset. In this case the appropriate transformation of the probability distribution has to be performed (for instance a normal distribution for log-transformed, log-normally distributed data), a procedure which could be cumbersome for some distribution functions.

Then, in order to have a good mathematical representation of the probability distribution, the cumulative distribution function (CDF) should be used, that it is directly related to rank-plots [550]. In addition, instead of graphical-based estimation methods, maximum likelihood estimation (MLE), which is not dependent on the graphical representation, is a superior analytical method [219, 411, 544, 262, 103]. Furthermore, the use of MLE allows to perform a statistical test of different proposed models [219, 103]. A rigorous quantification of goodness-of-fit to establish

relatedness to various distributions is critical for proper analysis of the empirical distributions.

1.3 Dynamical Analysis of Mathematical Models of Biochemical Interest

Once the graph properties of biochemical models have been introduced, we jump here to the questions relating to dynamical behaviour of biochemical systems. From a coarse-grained perspective determining which element interacts with which is sufficient to characterise the system as a whole. We have seen that many variables can be studied, which conduce us to an interesting knowledge about the biology of the organism. However in some cases, not only the network of interactions is known, but for a subset of the network, it is known quantitatively the terms of relationship for each link. When such a detailed information is available, the systems studied tend to be much smaller in size, from the order of tens to hundreds of nodes, not covering the complete organism. Systems modelled quantitatively refer generally to specific parts of the complete network, focusing on a biological process: for example the pathway of a given metabolism, a gene regulation module, etc.

At this point we have systems of relatively few items but with a much richer relationships between them. A first matter to take into consideration is the level of detail for the description of the interactions. A large variety mathematical formalisms have been used to model such systems, each of them accounting for different levels of description. Different methods to model dynamically biological systems have been compared [595, 120, 441].

Two main branches for modelling methods separate whether if state variables are continuous or discrete. Discrete models have been shown to be specially useful to establish the first hypotheses about the dynamics of biological networks of interactions (see Chapter 12 of the Kriete and Eils book [315] for successful examples). Some of them are Boolean networks [292, 555, 553, 293, 398, 528, 337, 608, 88] including cellular automata [160, 600, 132] and Bayesian networks [187] listing neural networks [167,

66]. Petri nets have also been used to model biological processes in general [430, 92, 438, 429, 86, 537, 603, 476] and specifically for metabolic pathways [462, 223, 199, 367, 583, 401, 488, 244], transcriptional regulation [139], gene regulation [366, 138, 400, 229, 397, 332, 137, 238] or signal transduction [367, 245, 325, 333, 331, 484]. Other modelling methods take continuous state variables but handle time as discrete like piecewise models [110], logic models [616, 615], pair-wise methods [24, 93], difference equations [552, 133], expression clustering [594, 569] or dose response equations based on weight matrices [593, 569].

Alternatively, state variables and time can be understood as continuous models. The majority of continuous models are formalised as differential equations (DEs). DE have a long successful record in the field of mathematical biology and the number of studies applying them is huge. Historic important models are the Hodgkin-Huxley model [256], the Lotka-Volterra equations [350, 348, 349, 580, 581, 215] or the Turing mechanism [565]. Examples of adequate books for the introduction of DEs in biological modelling are [551, 161, 399, 282, 308, 128]. Other interesting efforts are the collection of mathematical models of biological interest, stored in computer readable format and based on DEs. Some of the main examples are the BioModels Database [322] and the CellML repository [346] (see Section 1.3.2).

Many are the flavours of DEs. Once we are aware of their usefulness, I give here some hints about their formal definition. A DE is an expression of the relationship between functions and their derivatives.

$$y' = f(x, y) \tag{1.22}$$

Several notations are valid to express the same relation,

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x}, \tag{1.23}$$

the description of the change of state variables with respect to other independent variables. Most commonly the independent variable is time, then another possible notation raises:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \dot{y} \tag{1.24}$$

An idea important to take into account is the order of a DE, which is the order of the highest derivative of the unknown function. In the case the dependent variable is a function of a single independent variable, the DE is termed ordinary differential equation (ODE). If that function is linear, the system can be treated analytically but for the case of non-linear functions, generally the system needs to be solved numerically. ODEs have supplied successful stories on the modelling of a broad variety of biological phenomena. General overview can be found in [282] and a non-exhaustive and subjective list of particularly interesting examples is [389, 468, 107, 155, 193].

Discrete events embedded in continuous models of ODEs define a type of hybrid models commonly used in systems biology (a significant proportion of models from the BioModels Database contains events). Events can be dependent on time or other variables and can affect both state variables or parameters. A typical example is the cell division, when at a given point a single cell separates into two different ones.

Other types of DEs are partial differential equations (PDEs), in which, as opposed to ODEs, the dependent variable is a function of multiple independent variables. For example, for two independent variables, $y = u(x_1, x_2)$, the general form is,

$$F(x_1, x_2, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \frac{\partial^2 u}{\partial x_1 \partial x_1}, \frac{\partial^2 u}{\partial x_1 \partial x_2}, \frac{\partial^2 u}{\partial x_2 \partial x_2}) = 0. \qquad (1.25)$$

Both ODEs and PDEs can be classified as linear if the dependent variable and its derivatives appear to the power of one or non-linear otherwise. In fact, applied to the modelling of biochemical systems, power-law formalisms have been broadly used: the S-system formalism [500, 501, 502, 503, 505, 579, 504, 577, 530, 578, 562, 531, 572, 575, 354]. Equivalently, the number of applications using PDEs to the modelling of biological systems is immense. As representative works, it should be mentioned the works in pattern formation [565, 202, 376, 357, 5] or calcium signalling [207, 609]. Adequate deep knowledge of biological modelling using PDEs can be found in [399, 320].

Other types of DEs are differential algebraic equations (DAEs). DAEs are a type of DE where some of the dependent variables are not expressed

explicitly [436, 226]. The general form of a DAE is,

$$F(x, y, y') = 0, \tag{1.26}$$

or

$$\begin{cases} \frac{\mathrm{d}y}{\mathrm{d}x} = f(x, y, z) \\ g(x, y, z) = 0 \end{cases} . \tag{1.27}$$

DAEs have been used for modelling constrained dynamical systems, also in the field of molecular biology [420, 91, 113, 51].

DEs can hold special types of functions being therefore classified accordingly. Piecewise functions are functions which definition changes depending on the value of the independent variable. Piecewise differential equations (PWDE) [117] have been used to model several biological phenomena in which the system displays different behaviour after a defined threshold, like for example neuron models [372, 559, 561, 560] or some gene regulatory networks [210, 209, 523, 124, 125, 123, 84, 478, 196, 228, 126, 354]. Other works deal with sigmoid continuous threshold response functions [272]. In this case, the modelling framework provides with some advantages with respect to discontinuous functions [125], i.e., sigmoid functions are more realistic to gene regulation. Still considerable assumptions, that might be far from biological reality, need to be taken during the analysis of the model, like linear regulatory terms (see assumption $\mathcal{A}$ in [272]).

Another type of DEs used to model biological systems is delay differential equations (DDEs) [156, 56]. DDEs consist on DEs in which the derivative of some parts of the dependent variable are a function of the not immediately past system state, being,

$$y' = f(t, y, y_t) \tag{1.28}$$

where $y_t = \{y(\tau) : \tau \leq t\}$ represents the solution of y at a certain moment of the past. Solving DDEs requires knowing not only the current state of the system but also the state at certain previous times. Several examples of biological phenomena explained by DDEs can be found in [218, 507, 69, 112, 258, 402, 521, 31, 59, 522, 329, 334, 617, 224, 301].

Finally, a very special type of DEs are stochastic differential equations (SDEs). SDEs [421, 192] are characterised by a stochastic processs at

some of their terms. The general form for a system of SDEs is

$$\frac{\mathrm{d}y}{\mathrm{d}x} = g(y) + \sum_{m=1}^{n} s_m(y)\eta_m(x), \tag{1.29}$$

where y represents the dependent variable, g and s are arbitrary functions and η_m different types of random fluctuations. SDEs are essential to model biological systems [369, 23, 221, 222, 497, 470, 599, 485, 573, 597] specially when the deterministic formalism can not be assumed anymore, as for example, for regulatory networks whose nodes are found in low concentration.

Other types of DEs, like complex DEs, for which the solutions are complex functions, have been less popular to model biological systems.

Worth to note is that the types of DEs exposed here are not exclusive. As an example, delayed stochastic DEs have been used to model a variety of cellular networks [46, 482, 556, 621].

At the end of the day, once a mathematical formalism is chosen, the ultimate target of biological modelling, as for other disciplines as engineering, is the ability to design reliable and robust (biological) systems *in silico* prior to fabrication [47, 109]. Mathematical modelling is the tool to organise and predict the behaviour of biological systems. The choice of the modelling method should be done carefully, evaluating the type of questions we are interested in, because the results of the analysis may be different depending on the approach used [441]. DEs have been the approach selected for the current work. In this direction, some of the formal tools available to the modelling community are presented in the next section. Using such tools, as for example, sensitivity analysis, model calibration or optimal experimental design, those mathematical models of biochemical interest can be characterised, classified and understood.

1.3.1 Analysis of Analysis

In the previous section, I have focused on the use of different flavours of differential equations for the study of biochemical phenomena. Thanks to

differential equations we can predict the state of the system along time or space. We can determine the system stability, the effect of missing nodes from the system, different response depending on the initial conditions or more complex modifications of the system. Operation research can be applied to models of differential equations to solve questions as which model from a set of plausible ones is able to reproduce the experimental observations, which are the experiments necessary to validate our model or which reaction should be modified to obtain the maximum yield of a product. Rational strategies to answer those questions are described in the following sections.

While differential equations can be treated analytically, we have concentrated on the numerical approaches. Within the numerical approach, several types of analyses can be distinguished if the model dynamics are considered explicitly. Structural analysis like flux balance or elementary mode analysis do not centre its interests on the temporal dynamics but on the derived effects from the network connectivity [248, 315]. Alternatively, other methods work directly with the temporal evolution of the system variables. Some of them will be presented on the following sections. Others like frequency analysis, bifurcation analysis or stability of the system dynamics have been described elsewhere [270, 457, 566, 541]. In accordance to the results section, I introduce here some of the main computational tools used for numerical analysis of differential equations.

1.3.1.1 Model Integration

Numerical integration is a method to solve the direct problem: given a system of differential equations, it determines the system state at a given time point. The amount of research is huge on the field of ODE numerical integration. An adequate introduction to the topic is Chapter 7 from the book of R. Schwartz [512], I give here a succinct overview of the subject.

The most common technique to tackle the problem are finite difference schemes, which consist on the integration of the equations by summing successively over approximations at short time steps,

$$y_{t+1} = f(y_t, y_{t-1}, ...). \tag{1.30}$$

The simplest method is the forward Euler method consisting on the locally valid approximation,

$$y_{t+1} = y_n + \Delta t f(y_n), \tag{1.31}$$

for the generic system,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y). \tag{1.32}$$

An example of a more robust strategy is the Runge-Kutta methods that compute a series of intermediate middle points and uses a combination of them to determine the state of the system at the next time step. Four-order Runge-Kutta methods are a good compromise on stability and computational cost and they are in practice one of the most extended methods used in computational applications. Finally, I want to mention adaptative methods, like the Adams-Moulton, that can vary the step size accordingly to the stiffness of problem. In the computational biologists community, popular implementations of numerical methods for solving initial value problems are, among others, `LSODE` [253] and `CVODE` [104, 254].

1.3.1.2 Stochastic Simulation

Two assumptions necessary for models of ODEs are not always true for biological systems. Biological molecules are discrete and if the number of molecules modelled are below few hundreds, the approximation of the system by a continuous concentration is an oversimplification. Another accepted fact as true is the deterministic regime of the nature. Reactions occur thanks to random fluctuations and noise has to be taken into account for small systems.

Therefore if our system, or parts of it, behave out of these bounds, stochastic differential equations (SDE) are required for modelling the system [458]. System variables are not anymore concentration of the different species, c_i, but the number of molecules, $|y_i|$. The update of the system is then probabilistic. Given the system $\mathcal{S} = \{|y_1|, |y_2|, ..., |y_n|\}$, the state at time $t + \mathrm{d}t$ will be [205],

$$P(\mathcal{S}, t + \mathrm{d}t \mid \mathcal{S}, t) = a_i \mathrm{d}t + o(\mathrm{d}t). \tag{1.33}$$

being i a given reaction of the system and a_i the result of a mesoscopic reaction constant (c_i) and the number of particles.

$$a_i = h_i(y, c_i). \tag{1.34}$$

The summand $o(\mathrm{d}t)$ denotes for terms that can be dropped for sufficiently small $\mathrm{d}t$,

$$\frac{o(\mathrm{d}t)}{\mathrm{d}t} \to 0 \quad \text{as} \quad \mathrm{d}t \to 0. \tag{1.35}$$

That is to say, the variables of the system are now transition probabilities and the stochastic kinetic processes can be treated with the master equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}P(y_0, t_0, y, t) = \sum_{i=1}^{\nu} \Big(h_i(y - S^{(i)}, c_i)P(y_0, t_0, y - S^{(i)}, t) - h_i(y, c_i)P(y_0, t_0, y, t) \Big), \tag{1.36}$$

where ν is the total number of reactions and $S^{(i)}$ the ith column of the stoichiometry matrix.

Several algorithms have been implemented to solve the master equation, two of them implemented by Gillespie in the late 70s, the direct method and the first reaction method [204]. Due to the computational cost of those algorithms, other algorithms have been developed much faster with moderate loss of accuracy [201, 206]. Stochastic modelling has been widely applied to biological systems, reviewed in [597]. Finally I would like to point out to the book of Darren J. Wilkinson [599] as a complete introduction to the rich field of stochastic modelling of biological systems.

1.3.1.3 Sensitivity Analysis

The analysis of the system sensitivity (SA) consists on the evaluation of the variation of the output of a system with respect to some source of variation. This source of variation can be a change in the parameter values, initial conditions or a change in the model state variables.

In the context of SA applied to the model parameters, two very different levels of analysis can be distinguished, either local or global. For the local analysis, given the system's dynamics described by

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = f(\boldsymbol{y}(t), \boldsymbol{y}_0, t; \boldsymbol{\theta}), \tag{1.37}$$

where $\boldsymbol{y}$ is a vector of n state variables, $\boldsymbol{y}_0 = \boldsymbol{y}(t_0)$ is a vector of the initial conditions, $\boldsymbol{\theta}$ the vector of m model parameters and $t \geq t_0$, generally $t_0 = 0$. Then, SA can be defined as,

$$\boldsymbol{s}(t) = \frac{\partial \boldsymbol{y}(t)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial y_1(t)}{\partial \theta_1} & \frac{\partial y_1(t)}{\partial \theta_2} & \cdots & \frac{\partial y_1(t)}{\partial \theta_m} \\ \frac{\partial y_2(t)}{\partial \theta_1} & \frac{\partial y_2(t)}{\partial \theta_2} & \cdots & \frac{\partial y_2(t)}{\partial \theta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_n(t)}{\partial \theta_1} & \frac{\partial y_n(t)}{\partial \theta_2} & \cdots & \frac{\partial y_n(t)}{\partial \theta_m} \end{pmatrix}. \tag{1.38}$$

Similarly, for $1 \leq i \leq n$, $1 \leq j \leq m$,

$$s_{i,j}(t) = \frac{\partial y_i(t)}{\partial \theta_j}, \tag{1.39}$$

which is usually called unnormalised sensitivity. The normalised sensitivity,

$$S_{i,j}(t) = \frac{\theta_j}{y_i(t)} \frac{\partial y_i(t)}{\partial \theta_j}, \tag{1.40}$$

may confront numerical problems when the value of y_i gets close to zero. Numerically, an indirect method [604] to calculate the sensitivity is the use of the centred difference approximation,

$$s_{i,j}(t) = \lim_{h \to \infty} \frac{y_i(t; \theta_j + h) - y_i(t; \theta_j)}{h} \approx \frac{y_i(t; \theta_j + \Delta\theta_j) - y_i(t; \theta_j - \Delta\theta_j)}{2\Delta\theta_j}, \tag{1.41}$$

being commonly $\Delta\theta_j = 0.001\ \theta_j$ [626]. Computationally the sensitivity of the system with respect to the parameters is evaluated at $1 \leq k \leq l$ discrete points along the integration time and several normalisation strategies are available to render a specific value of sensitivity of the system with respect to a given parameter [538, 614, 626]. Alternatively to the finite difference method, other methods to determine the sensitivity of the system are the so called direct methods, which determine

$\boldsymbol{y}(t)$ and $\boldsymbol{s}(t)$ simultaneously [136, 147, 328, 516]. Other types of alternative local approaches to calculate numerically the sensitivity of a system [548, 270, 338] have been used to study biological networks. Sensitivity analysis has been applied to many different biological systems [538, 264], providing with knowledge about the dynamics, in many cases hard to find by intuition. Also, sensitivity analysis can be useful for model reduction as the parameters that affect weakly the system behaviour can be bypassed [130, 342].

Within the framework of metabolic reaction networks (enzymes and metabolites) in steady state, metabolic control analysis (MCA) can be understood as an extension of sensitivity analysis. MCA consists on a mathematical framework for the description of the relative changes of system variables (as steady-state concentrations, reaction fluxes) with respect to perturbations on the system parameters [170, 248, 171, 288, 596, 259]. MCA was firstly developed from initial ideas of Higgins [250] by two independent groups in the 1970s: Kacser and Burns [287] and Heinrich and Rapoport [247, 246]. A central measurement of MCA is the control coefficient, defined equivalently for both substrate concentration (S) and fluxes (J):

$$C_{v_i}^{c_j} = \left(\frac{A_i}{S_j}\frac{\partial S_j}{\partial A_i}\right) / \left(\frac{A_i}{v_i}\frac{\partial v_i}{\partial A_i}\right) \tag{1.42}$$

$$C_{v_i}^{J_j} = \left(\frac{A_i}{J_j}\frac{\partial J_j}{\partial A_i}\right) / \left(\frac{A_i}{v_i}\frac{\partial v_i}{\partial A_i}\right) \tag{1.43}$$

being v the reaction rate and A the activity of a given enzyme, which can be related to its concentration. The summation theorem is a systemic property of control coefficients,

$$\sum_i C_{v_i}^{J_j} = 1 \tag{1.44}$$

and

$$\sum_i C_{v_i}^{c_j} = 0. \tag{1.45}$$

Finally, another importat variable is the enzyme elasticity,

$$\epsilon_{S_i}^{v_j} = \frac{S_i}{v_j}\frac{\partial v_j}{\partial S_i} \tag{1.46}$$

which define the degree of change of reaction rate v_j with respect to substrate concentration S_i. Application of MCA approach has provided insights on several biochemical models [311, 108, 257, 82, 194]. Interestingly, an extension of MCA is frequency response analysis to perturbations [270, 457]. Moreover, bifurcation analysis can be indicated as a tool to examine qualitative changes on the behaviour of the system steady state with respect to changes in the model parameters [566, 567, 16, 588, 558, 154, 490, 251, 362].

Local sensitivity analysis provides a measure of the system behaviour at the closed neighbourhood of the nominal values. However, the sensitivity might be, generally, very different if calculated at a different parameter points. For that reason, if we lack a high knowledge about the parameter values, a global sensitivity analysis is recommended [99, 356]. Global sensitivity analysis consists on the use of statistical tools to address the behaviour of a system over a wide range of parameter values [271]. A first simple approach would be the Morris method [394] (used in [613]) which consists on taking the mean value of $s_{i,j}(t)$ over several sampled points of θ_j. Other more elaborate methods weigh differently the sampled parameter points, using for example, the Boltzmann distribution [60].

More commonly, a broad distribution of parameter values is sampled and the output of the system is analysed. Methods based on this strategy are random sampling—high dimensional model representation (RS-HDMR) [173], multi-parametric global sensitivity analysis [96, 623, 364, 94, 610], partial rank correlation coefficient [620, 364], Sobol's method [525, 310], Fourier amplitude sensitivity test (FAST) [111, 506, 309, 373, 364] and its related Walsh amplitude sensitivity procedure (WASP) [437]. An efficient sampling procedure for high dimensional systems is the Latin hypercube sampling used in many of the previous works cited above. Furthermore, several global methods have been compared using the same signal transduction pathway [620]. Other forms of global sensitivity analysis, like the information-based approach proposed by Lüdtke *et al.* [351] results very useful given the large number of species and parameters commonly found in models of biochemical systems.

Also, I would like to point out to the reviews from Turányi [564], Frey and Patil [186] and Saltelli [496, 495] for a complete description on the

topic of sensitivity analysis.

Finally I would like to emphasise the relationship between sensitivity and robustness. Robustness, contrary to sensitivity, is the system property to maintain its function under internal or external perturbations [305]. Several indexes have been proposed to account mathematically the robustness of a system [595, 89, 305, 518]. Indeed, many biological systems have been described as robust[8], [44, 53, 185, 582, 584, 393, 68, 539, 153, 625, 18, 518] among many others. Robustness and model parameter sensitivity can also be used as model validation [178].

1.3.1.4 Model Calibration

Once we have a mathematical model describing the dynamics of a system, we might wonder how far is our model from reproducing the observed behaviour of the system. In principle, one of our goals is the establishment of the simpler model that most accurately describes the phenomenon we are interested in. Initial steps are the choice of the mathematical formalism we are going to use (see Section 1.3) and the definition of the interactions among the network components. Often within the frame of systems biology, information is available about which components of the system interact with each other and to a certain level the qualitative type of relationship, e.g. activations, inhibitions, complex formation, etc. If knowledge about model parameters and initial conditions is available, the inference of the dynamics of the system is called *direct problem* and consists on the integration of the system or the solution of the equations (see sections 1.3.1.1 and 1.3.1.2). The information that most commonly is missing, however, is the quantitative strength of the interactions: the numerical values of the parameters that define the interactions. We face then the *inverse problem*. A prevalent approach is to calibrate the model in order to reproduce the observed behaviour, that is to say, to modify the model parameter values in such a way that the computed output match as best as possible the measured variables of the system. Model calibration could be one of the most challenging tasks in systems biology

[8] Do not mistake robustness with inconsistency [595]. While the first refers to an intrinsic property of the system, the second refers to a rather wrong balance between the model level of description and the sampled data available.

[241]. Obviously, when modifying the parameter values, the model will exhibit different behaviour, however the flexibility of the response will be within the bounds imposed by the fixed assumed structure of the model [595]. Certain dynamics will not be reached by changing the parameter values but changing the intrinsic structure of the model: model selection is a major challenge in systems biology. Automated symbolic reverse engineering of non linear systems has been demonstrated [73]. However, when we are confident of the specified rules of interaction among chemical species, model calibration can play an decisive role in model validation.

In some cases, the structure of the model itself allows to perform a relatively easy model calibration: for example S-systems [500, 501, 502] can be decoupled and non-linear terms can be linearised making possible of a much faster search on the parameter space [303, 576, 563, 197, 574, 575]. For most of these cases, data points can also be approximated to more smooth functions and those functions become surrogate models, bypassing consequently the computationally costly step of model integration. When such strategies cannot be applied, true for a large amount of non-linear dynamical models, the parameter space exploration using optimisation algorithms becomes much harder.

For this purpose a function called objective function is defined which accounts for the goodness of the model with respect to the targeted behaviour. Different types of objective functions can be introduced, depending on the interests one wants to meet, in fact, multiple objective functions can be optimised simultaneously [240]. If we assume the residuals are normally distributed, independent and homoscedastic, then the maximum likelihood criterion is equivalent to the sum of the least squares. The log-likelihood form,

$$\mathcal{L} = \frac{l}{2} ln(2\pi) + \frac{1}{2} \sum_{i=1}^{l} \left[ln(\sigma_i^2) + \frac{\left(\tilde{y}_i - y_i(\boldsymbol{\theta})\right)^2}{\sigma_i^2} \right], \tag{1.47}$$

can be reduced to the objective function,

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{l} \frac{\left(\tilde{y}_i - y_i(\boldsymbol{\theta})\right)^2}{\sigma_i^2}, \tag{1.48}$$

where l is the total number of *constrains*, $\tilde{y}_i \in \mathcal{Y}$. Constrains are mapped

to variables through the observation function, h,

$$\tilde{\boldsymbol{y}} = h\big(\boldsymbol{y}(\boldsymbol{\theta})\big). \tag{1.49}$$

From here on, a *constrain* is assumed to be the value of any of the system variables at a given time. They can be obtained from experimental measurements ($\tilde{y}_i$) and they are compared to their corresponding values obtained from simulation ($y_i(\boldsymbol{\theta})$). σ_i^2 is the variance associated to $\tilde{y}_i$. Note that system variables do not hold necessarily the same number of constrains. In practice other several similar forms have been used as,

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{l}\frac{\big(\tilde{y}_i - y_i(\boldsymbol{\theta})\big)^2}{\sigma_i^2}, \tag{1.50}$$

or

$$J(\boldsymbol{\theta}) = \frac{1}{l}\sum_{i=1}^{l}\frac{\big(\tilde{y}_i - y_i(\boldsymbol{\theta})\big)^2}{\sigma_i^2}, \tag{1.51}$$

Therefore, depending on the parameter values, the value of $J(\boldsymbol{\theta})$ will vary on the parameter space (see Figure 1.3). I would like to remind that, in the case the initial conditions of the system ($\boldsymbol{y_0}$) are unknown, they can be considered as parameters to calibrate and therefore $\boldsymbol{y_0}$ may depend on $\boldsymbol{\theta}$. A huge number of strategies of largely different nature have been conducted to explore the objective landscape. In principle there is no a single method that outperforms all others for a representative variety of problems [379, 34]. The election of the most suitable method for the problem in hands is generally tricky and often left to a trial and error strategy.

Three main groups can be distinguished based on the way the parameter domain is explored: local search, global search and a third group designed as the hybrid approach which consists on a more or less elaborated mixture of both methods. Additionally, other alternative model calibration strategies have been applied to biochemical dynamical systems, like maximum likelihood estimation [80] or Kalman filters [547]. Others stress the importance of an initial phase of constrain data processing [455, 214] or the capability of model selection [339]. Although not universally applicable, other complex approaches following the *divide and conquer* spirit have been deployed [433] for the calibration of gene regulation models performing orders of magnitude more efficiently. Other less

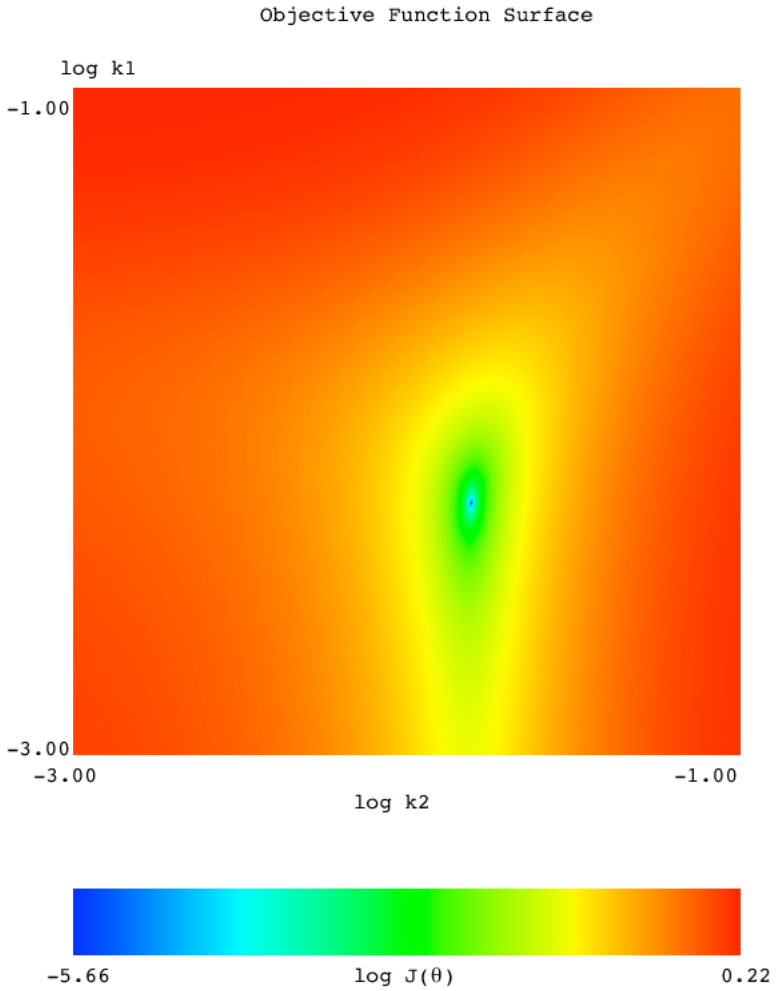

Figure 1.3: Objective function ($J(\boldsymbol{\theta})$) surface for a 2D parameter space. In this convex situation, it can be observed that the surface is rather flat with a deep and narrow basin of attraction at the nominal parameter values. All in all, these types of surfaces are relatively easy to explore given the existence of a sole minimum. In other cases (see Figure 1.4), surfaces are rough and the large abundance of spread basins of attraction makes model calibration a cumbersome task.

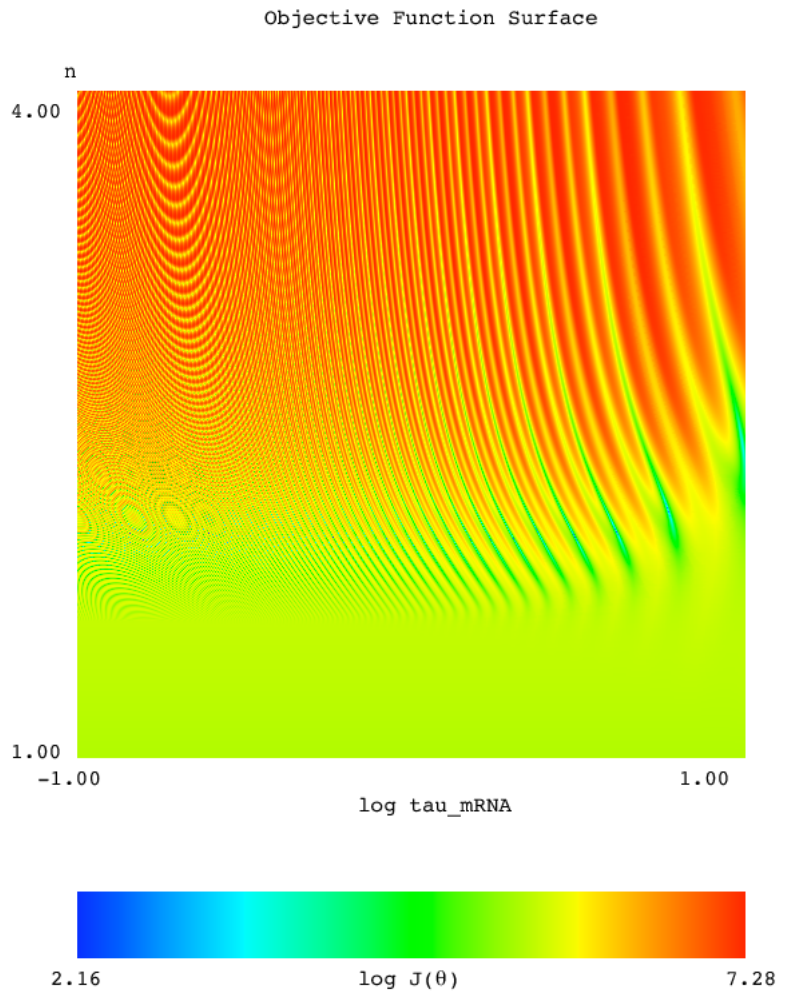

Figure 1.4: Objective function ($J(\boldsymbol{\theta})$) surface for a 2D parameter space. Contrary to the situation of Figure 1.3, here a multi-modal (non-convex) objective surface is presented. It is largely rough and full of local minima: observe the blue dots around the right lower third of the figure. Fair chances of finding the global minimum in such surfaces requires from the combination of local and global strategies and a large computational effort. Still the solution found is generally suboptimal.

popular methods have also been applied to biological systems; I suggest to consult the prolific review of Chou and Voit [97] for a comprehensive list of model calibration methods. General reviews about model calibration of biological systems can be found in [390, 25, 39]. More methods are compared in [198].

Worth mentioning is that the previously described methods can be applied to a wide range of modelling frameworks, including both deterministic and stochastic models [469, 557].

1.3.1.4.1 Local Search Methods Local search methods are iterative methods that evaluate at a given point the topological characteristics of the objective surface by, for example, calculating directly the gradient or approximating it. From this information, a new point in the parameter space is selected and a new step begins. The algorithm stops at different convergence criteria, like a very low value of the objective function, deprecative improvement of the objective function, minor movements in the parameter space, achieved maximum number of steps, etc. Some of the most common local search methods applied to biological models are based on Levenberg-Marquardt or Gaussian-Newton methods [475, 25].

1.3.1.4.2 Global Search Methods In the case the objective function surface presents many local minima, the application of a local search approach will render inconsistent results: starting from different points of the parameter space, the final parameter values will differ. In those cases, global search methods are recommended in order to surmount the problems derived from the non-convexity situation. Global methods can be classified as deterministic or heuristic [475] and several of the most popular heuristic ones have been compared in [34]. Generally the size and complexity of typical models of systems biology exceed the capabilities of deterministic methods and therefore I describe here some of the most popular methods used in the field:

- *Random or multi-start:* When local search methods fail to provide consistent results, a first approach to get an idea of the shape of the objective surface would be multiple shooting [70, 428, 38, 300],

that's it, to launch several instances of a local search method from different starting points of the parameter space. It constitutes a good compromise for parameter spaces of low dimension and analysis of the results can help the detection of the major optimal regions of the parameter space.

- *Metropolis-Hasting:* Metropolis-Hasting sampling [243] consists on the exploration of the surface randomly and accepting parameter space points depending on a probability calculated from the objective function. It has been applied to several biological problems [498]. Although it is a relatively cheap method, computationally speaking, its performance may be reduced for high dimensional functions or functions holding steep narrow local minima.

- *Simulated Annealing:* Simulated annealing [304, 98] is a metaheuristic global optimisation method consisting on the probabilistic choice of new states. The probability depends on the value of the function we want to minimise and on a parameter called *temperature* that is gradually decreased during the process. When temperature is set at high levels, broad parameter space can be searched. Then when the temperature is gently decreased, the lower regions of the objective surface are explored. Simulated annealing has been applied to the calibration of many biological models, [468, 467, 379, 390, 276, 275, 220] to cite a few.

- *Evolutionary algorithms:* Evolutionary algorithms is another metaheuristic approach for combinatorial problems. It consists on a population of instances that based on random movements explore the parameter space guided by the objective function. Some of the most commonly used are genetic algorithms, particle swarm optimisation or ant colonisation optimisation. These algorithms have been widely used for the calibration of dynamical systems of biochemical interest [379, 390, 281, 177, 284]. An special mention deserves the scatter search method [150, 151]. It has been reported [474, 284] that these methods have outperformed, both in accuracy and speed (several orders of magnitude indeed) other previously state of the art methods for several benchmarks within the biological modelling field.

Other global approaches, like exhaustive strategies [440] as the branch-and-bound method, have been successfully applied to biological models, specially to generalised mass action models.

1.3.1.4.3 Hybrid Methods: While local search methods converge fast, they suffer from non-consistency on rough surfaces. On the other hand, global methods are able to surmount the problem of multiple local minima but they are very slow finding the lower parts of the objective surface. Therefore a straightforward approach would be the combination of both methods [276, 275, 563, 628, 475, 177, 37, 26, 27, 255, 300]. A global method may be used initially to locate widely the most optimal regions and from there, a local search method would be triggered. This combination of approaches benefits from the virtues of both methods [40] and provides with very good results with minor complications (a transition rule should be defined for the change from the global to the local search phase.).

1.3.1.5 Identifiability

The previous sections, approaches for model sensitivity and model calibration have been presented. In a close relationship, including concepts of both topics, I introduce here parameter identifiability. Parameter identifiability refers to the ability to determine the numerical value of a model parameter from experimental data. Formal definitions can be found in Grewal and Glover [225]. A great review to bring into scene the problem of identifiability in biological models is [278]: it treats the key mathematical concepts in a very amenable manner and furthermore it contains enlightening examples of the forefront biological research. Intuitively, there are three main reasons for lack of identifiability [534]:

- *Degrees of freedom derived from the number of constrains and parameters.*
- *Parameter correlations.*
- *Noisy measurements.*

One of the major reasons for the failure of model calibration methods is the presence of flat regions in the objective function: when the algorithm arrives to a flat region, it is not capable to find a sensible direction to go further and it gets trapped there. Flat regions on the objective function can be expected if the system displays very low differences from the constrains (the experimental data) for variations in the parameter range [236]. Flat regions are associated with a problem of large number of degrees of freedom. Also if model parameters correlate, that's it, if different parameter combinations yield the same output (evaluated through the objective function), a lack of identifiability is also presented. It can be observed when the sensitivity of a parameter is a constant multiple of another [534]. From an alternative perspective, identifiability is not attainable when deviations in the parameter space do not have a great impact on the objective function [474]. In this direction it would be interesting the comparison of Figure 1.3 and 1.4 and observe the radically different behaviour of the objective function along the parameter space. Identifiability analysis, then, becomes a great complementary tool for model calibration as it helps on the interpretation of model calibration results. In fact, parameter identifiability should be studied prior to any model calibration to ensure that the parameter estimation is well-posed [474].

Identifiability analysis can be distinguised as two types: *a priori* and *a posteriori* identifiability. *A priori* identifiability integrates the first two sources of unidentifiability (named, degrees of freedom and parameter correlations) while *a posteriori* identifiability comprehends all three [534], including the observational noise. *A posteriori* identifiability alludes to the identifiability of the parameters given a finite set of constrains or experimental values $\mathcal{Y}$ while *a priori* analyses the model itself, without any specific set $\mathcal{Y}$. *A priori* identifiability evaluates whether the parameter for the mathematical model can be determined assuming that for all observables continuous and error-free data would be available [25]. Moreover identifiability can be classified as local or global if we talk about a specific region in the parameter space or its complete domain.

1.3.1.5.1 *A priori* Identifiability Although some techniques are available [87, 30, 163, 57, 381, 487, 619] and they have been applied to

biological models [30, 36, 477], *a priori* global identifiability is a hard task in non-linear dynamical models [475, 452, 459]. When possible, *a priori* global identifiability should be tested [278]. Important insights about the structure of the model itself will be uncovered. On the other hand, if we work at a precise point of the parameter space, for the local analysis, simply the correlation of the sensitivities (see Equation 1.39) is calculated [618]. Any correlation of +1 or -1 makes those parameters *a priori* unidentifiable. One of them should be fixed and the identifiability analysis should be repeated once again until no unidentifiable parameter is detected.

1.3.1.5.2 *A posteriori* Identifiability When we have access to temporal experimental data $\mathcal{Y}$ for some of the state variables of a model, we can perform *a posteriori* identifiability for that data set. Assuming Gaussian noise on the constrains, a local approach is based on the Fisher information matrix (FIM). It is very similarly to the *a priori* approach, but restricted to experimental data points.

$$\text{FIM} = \sum_{i=1}^{l} \left(\frac{\partial y_i}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}} Q_i \left(\frac{\partial y_i}{\partial \boldsymbol{\theta}}\right), \tag{1.52}$$

where $l = |\mathcal{Y}|$ and y_i refers to the time and state variable corresponding to $\tilde{y}_i \in \mathcal{Y}$. $Q = M^{-1}$, being M the measurement error covariance matrix. When parameter correlations are present (larger than 0.99, for example) the FIM becomes singular and the model is not identifiable [475, 474]. Another important concept is the correlation matrix, R,

$$R_{i,j} = \begin{cases} \frac{C_{i,j}}{\sqrt{C_{ii}\, C_{jj}}} & \text{if } i \neq j, \\ 1 & \text{if } i = j, \end{cases} \qquad \text{for } i, j = 1, ..., |\boldsymbol{\theta}|, \tag{1.53}$$

obtained from the parameter estimation covariance matrix, C, which is approximated as the inverse of the FIM.

Finally, in the neighbourhood of the estimated parameters $\hat{\boldsymbol{\theta}}$, a linear approximation can be used to measure the quality of the estimates. Confidence intervals can be calculated through the FIM [475] as well as from other alternative approximations [474, 534] like the second derivative (the

Hessian matrix) of $J(\boldsymbol{\theta})$ [26] or more robust methods as jackknife or bootstrapping [313, 33, 283]. Systematic exploration of the objective function is also possible [474], however this approach is computationally too expensive, specially for typical models of biological systems. Confidence intervals around the estimated parameter $\hat{\boldsymbol{\theta}}$ can be calculated thanks to the Cramér-Rao inequality [344], which states that:

$$\sigma^2_{\hat{\theta}_i} \geq \text{FIM}^{-1}_{ii}, \quad \text{for } i = 1, ..., |\boldsymbol{\theta}|. \tag{1.54}$$

Therefore, the true parameter value $\boldsymbol{\theta}^*$ lays in the interval [283],

$$\hat{\theta}_i - \sigma_i t^{\nu}_{\alpha/2} \leq \theta^*_i \leq \hat{\theta}_i + \sigma_i t^{\nu}_{\alpha/2} \tag{1.55}$$

where $t^{\nu}_{\alpha/2}$ is the t-Student statistic for ν degrees of freedom and a confidence level of $(1-\alpha)$. These equations should be be taken with a grain of salt as we are assuming a linear approximation and Equation 1.54 is only a lower bound to the variance [283]. Revealing graphical interpretations of parameter confidence intervals can be found in [26, 25].

It is very important we do not start a model calibration until no correlations are present, otherwise we face very poor chances of successful parameter estimation. Not only model calibration methods will fail, but the obtained results will be meaningless, essentially random numbers. Identifiability analysis have been applied to different biological systems [274, 76, 618, 75, 313, 191, 614, 236, 452], including power-law models [534] and *in silico* models [459]. The application of identifiability analysis is strongly recommended for models of biological interest, for which large emphasis is put on the biological interpretation of the estimated parameter values [26]. Ashyraliyev *et al.* have studied the *a posteriori* identifiability of the *Drosophila* gap gene circuit [26] uncovering significant problems: due to model parameter correlations, none of the parameters determined presented reasonable accuracy, a problem that can not be always relieved with better experimental data sets. Gadkar *et al.* [191] conclude that some models, given their lack of identifiability, require from the direct measurement of certain key parameters (reaction rates, for example) to capture the dynamics of the experimental system.

Apart from the FIM-based approach to identifiability, other alternative methods are presented and compared in [452].

Although not strictly *global* approaches of *a posteriori* identifiability, I would like to stress the importance of some research works on the study of *a posteriori* identifiability along different parameter points of a bound parameter space. Examples of such an important approach is the work of Balsa-Canto *et al.* [36] or Ashyraliyev *et al.* [26]. Identifiability is derived from averaged values of sensitivity from a Latin hypercube sampling or FIM values from optimal model calibration results respectively. It is important identifiability is evaluated at different representative parameter locations due to the generally large roughness of the objective function.

Finally, worth noting are two excellent books for a deeper understanding on identifiability methods [590, 344].

1.3.1.6 Optimal Experimental Design

In the previous section I have presented a pervasive problem on biochemical modelling: parameter unidentifiability. At the worst cases, even using ideal data (both in quantity and quality, i.e., sampling within the order of simulation time step for all the state variables and noise free data respectively, a situation largely out of reach of current experimental settings), specific model parameters cannot be constrained tightly. Not in the sense to solve it completely but in the spirit of making the problem less severe, the techniques of optimal experimental design (OED) appear. An intuitive working hypothesis of OED is the following: *do constrains ($\tilde{y}_i \in \mathcal{Y}$) have a different influence on the model calibration problem?* Accepting the hypothesis as true and understanding the model calibration problem as a function minimisation (the objective function, $J(\boldsymbol{\theta})$), OED searches for the constrains that transform J in such a way that it results more facile for the algorithms to explore it and eventually arrive to the global minimum. As expressed in [35], OED aims to define the measurements *with the maximum amount and/or quality of information for the subsequent model calibration.* A work worth consulting that present the problem of OED very clearly is [41], an excellent broad review introducing the OED in the context of systems biology is [314] and deep insights on OED could be found at the A. C. Atkinson book [29].

Within this framework, diverse forms are available to discriminate con-

strains in such a way. Mainly OED is based on parameter identifiability results and it can be broadly distinguished as in local and global designs.

1.3.1.6.1 Local OED The most suitable prospective experiment would be the one that maximises a given criterion. Using the FIM, several optimal criteria can be defined [475]:

- *A-optimal design criterion:* Defined as,

$$\hat{\mathcal{Y}} = \arg\min_{\mathcal{Y}} \ \mathrm{tr}\big(\mathrm{FIM}^{-1}\big), \tag{1.56}$$

  which can be interpreted as which set of constrains, $\mathcal{Y}$, provides with the lowest trace of the inverse of the FIM. It can be interpreted as an effort to minimise the average variance of the parameter estimates.

- *Modified A-optimal design criterion:* Equivalent to the A-optimal design criterion, it is defined as the maximisation of the trace of the FIM:

$$\hat{\mathcal{Y}} = \arg\max_{\mathcal{Y}} \ \mathrm{tr}\big(\mathrm{FIM}\big). \tag{1.57}$$

- *D-optimal design criterion:* A geometric correspondent match of the arithmetic strategy of the modified A-optimal design criterion is the D criterion, which maximises the determinant of the FIM:

$$\hat{\mathcal{Y}} = \arg\max_{\mathcal{Y}} \ \det\big(\mathrm{FIM}\big). \tag{1.58}$$

  That implies minimising the overall volume of the estimated parameters' variance [35]. Kutalik *et al.* [318] elaborated a formal procedure to optimise the constrain that better improves the D-optimal criterion.

- *E-optimal design criterion:* Defined as,

$$\hat{\mathcal{Y}} = \arg\max_{\mathcal{Y}} \ \lambda_{min}\big(\mathrm{FIM}\big), \tag{1.59}$$

  where $\lambda_{min}\big(\mathrm{FIM}\big)$ is the lowest eigenvalue of the FIM, the E-optimal design criterion is understood as an effort to minimise the variance on the widest principal component of the FIM.

- *Modified E-optimal design criterion:* Instead of maximising the lowest eigenvalue as for the E criterion, in this case the goal is to limit the extreme eigenvalues,

$$\hat{\mathcal{Y}} = \arg\min_{\mathcal{Y}} \frac{\lambda_{max}(\text{FIM})}{\lambda_{min}(\text{FIM})}. \tag{1.60}$$

Geometrical interpretation of the OED criteria can be found in [28]. Several of these criteria have been applied to various biological experimental setups [41, 162, 613] with extraordinary results [38]. Additionally other FIM-based criteria have been applied to signal transduction pathways [83]. Finally, robust extensions of the local FIM-based criteria to the neighbourhood of a parameter point have been developed [28, 176] and applied to a signal transduction pathway [77, 613].

1.3.1.6.2 Global OED Local OED approaches assume that we know with precision the parameter values or at least that our estimates are very close to the real values. But for most of the cases in biological modelling, this is not true and estimates of the parameter values are largely broad. For example, the calculation of the FIM-based criteria could be radically different at different local minima of the parameter space. Therefore global OED approaches would be more adequate when poor information is available about the numerical values of the model parameters. Although computational cost would be, in general, larger, predictions would enjoy greater robustness, against experimental measurement errors, for example. Balsa-Canto *et al.* [35] presented an application of global OED to a signal transduction model, taking as a basis the concepts of FIM-based optimal criteria. Starting from a Monte Carlo sampling of the fitness landscape, the shape of the results is analysed. A PCA is performed on the 0.05-0.95 interquantile, generating several criteria suitable for OED:

- *Volume of the hyper-ellipsoide:* Taken as the product of the semi-axes.
- *Maximum eccentricity:* Taken as the ratio of the largest semi-major axis divided by the smallest semi-minor axis.

- *Distance to nominal values:* From the parameter nominal value, the distance to the mean value of the cloud of points is determined.

The volume variable can be understood as a global equivalent to the D-optimal criterion (Equation 1.58), while the eccentricity is a simile to the modified E criterion (Equation 1.60). Other methods take as criterion the data that provide a smallest global parameter sensitivity, Feng and Rabitz [174] presented an elaborated protocol in which a genetic algorithm determines which experiment would be capable to reduce furthest the parameter global sensitivity. Despite of its large computational cost, its method is able to propose sensible experiments without the requirement of any knowledge about the precise value of the model parameters. Moreover, the same authors presented two years later an improved approach [175] taking into account the extremes of the parameter distribution: parameter distribution analyses after model calibration could be a computer expensive approach, but a very appropriate criterion to direct global OED. Similar ideas have been used to assess indexes of global sensitivity on MCA models [491].

Alternative protocols for global OED have been proposed recently [241], unfortunately they have been developed for discrete time models (difference equations) limiting therefore its direct deployment on DEs models.

Both local and global OED can be included into iterative protocols along with model calibration [190, 38] in an attempt to determine with certainty the model parameter values. Finally, I would like to note that OED can be not only be applied to model calibration, but also to model selection: [314, 549] and references therein.

1.3.2 Information Storage and Sharing

Up to here, I have surveyed the broad range of methods available to analyse the dynamical responses of biological models. Suppose that you have a large interest on a specific pathway and you want to perform some of the analyses described above. For example, you would like to know which of the kinetic rates is the most sensitive of the system. How should

we proceed? Well, now I will shortly introduce here some of the main computational pieces necessary to carry out the presented approaches.

First of all we need a computer format to store the biological model. Within the experimental biology community, biochemical pathways are generally shared using cartoons of arrows and names. However if we want to work quantitatively with the system, we require from computers and therefore a format to store it. Following this direction whatever computer format would be possible, from a text file to a Fortran code file. But we should choose a format, though, that could be able of taking profit of the plausible befits offered by them. By far the most suitable one is SBML [268], standing for Systems Biology Markup Language. SBML is a computer-readable format based on XML for representing models of biological processes, being the *de facto* standard format in computational biology [144] with over 180 software tools (July, 2010) supporting it. SBML enables us to use the model with different tools without the need to re-write it, which is largely convenient, avoiding thus the incorporation of errors. Moreover, a SBML model can be shared among different scientists using different tools or programming languages, or even sharing it publicly, in repositories or journals. Further, models can become independent of the software they were created with, developed or analysed. Extensive resources of information are available about SBML from its portal, `http://sbml.org/Main_Page`. A similar effort to SBML is the format CellML [345] and tools exist (`CellML2SBML` [508]) that automatically convert both formats.

Concentrically to SBML as the nuclear seed, other computational resources have been developed. One of the most relevant is concerning the visualisation of the models. The Systems Biology Graphical Notation (SBGN) [324] provides with a way to represent biological pathways in an unambiguous way. Another important aspect of biochemical modelling, aside from the model itself, is quantitative experimental data. As we have seen in Section 1.3.1.4, quantitative data for the behaviour of the system is of crucial importance for model validation. Computational standards to store and share such data have been recently developed within the SBML community, the Systems Biology Results Markup Language (SBRML) [116]. Not only experimental data, but results of a simulation with reference to a SBML model can be specified. Comple-

mentary efforts, specialised on microarray data sets have been developed too [489].

Another of the most important pieces of the SBML world are model repositories publicly accessible. BioModels Database and JWS Online Model Database store about 250 and 90 models respectively (in July 2010). Models of disparate biological interest (metabolism, signal transduction cascades, genetic networks, etc.) can be downloaded and simulated on-line.

Finally we want to conclude this section with a reference to MIRIAM (Minimum Information Required in the Annotation of Models) [323] which can be understood as a guide for scientists to develop and share computational models of biological interest.

1.3.3 Software Tools for Models

The amount of software developed for the field of systems biology is large and heterogeneous. It is out of the scope of this introduction to review carefully in detail the complete broad range of examples. As I said previously, more than 180 tools are available, just for models written in SBML. A tentative review of the tools available for kinetic modelling of biochemical models, where some of the major programs are evaluated, is a good source of primary information [12]. Instead I will give here a subjective overview of the most useful tools available in the field. Another source of information with short description about the capabilities of the SBML-compatible applications can be found in the SBML portal, specifically in `http://tinyurl.com/2c44sq6`.

The very first step would be the creation of the model. We need a tool to build a model and preferably to visualise it. Major tools used for this purpose are `CellDesigner` [189] and `COPASI` [263, 378] among many others (`http://tinyurl.com/32eenag`). A very helpful tool during the creation of SBML models is `SBMLsqueezer` [145] which helps on the definition within the model of the mathematical equations of the kinetic laws. If different single models need to be merged, modular composition of different SBML models can be assisted by `SBMLmerge` [511]. With-

out the need of specific reaction kinetics, qualitative analysis of SBML models can be conducted with or SQUAD [135]. Generally, a graphical interface exists for these tools and if our model is based simply on ODEs, integrating the system and generating the dynamics is nicely accessible for scientists with no high skills on programming. Narrator [358] offers not only a graphical interface to build a model but also provides with techniques on information processing to facilitate the development of the model. Other tools providing deterministic and stochastic simulators are also available. If we are interested on stochastic simulations, I would recommend DIZZY [456], STOCKS [297], Cain or the efficient CellMC [85] as a good starting point. An special mention is required for spatial simulators, applications that take into account the 3D nature of cells or multi-cellular entities. Biological processes in which reaction/diffusion phenomena are key factors can be modelled using among others Virtual Cell [392], CompuCell [273] and SmartCell [14] which have been shown to be the most matured. Not compatible with SBML is Smoldyn [15], which is specially suitable for modelling mesoscopic aspects of cellular biology.

Other resources like the Systems Biology Workbench (SBW) [499] or the Systems Biology Toolbox for MATLAB [509], they are more than a single application, being true environments where a combination of different applications converge to provide the user with a diverse bunch of analyses, from deterministic to stochastic simulations, different types of sensitivity analysis, model calibration, etc.

Specifically for sensitivity analysis, the number of tools is large. The most complete ones are without doubt SBML-SAT [626] and SensSB [473], but there exist others as well as BioSens. For the related field of metabolic analysis, two main tools exist integrated in the MATLAB environment, the COBRA Toolbox [52] and CellNetAnalyzer [306].

The important analysis of model calibration has been included in a very large number of applications. Some of the ones I consider outstanding are COPASI, SBML-PET [624] and DOTcvpSB [255]. All three incorporate many different strategies for function minimisation, from local to global methods including hybrid algorithms. Special mention requires software dedicated to parameter inference for stochastic models, like CaliBayes

[95] which is capable of distributing the large computational cost derived from such an expensive task.

Although identifiability analysis has been less popular among systems biology scientists, efforts to provide software for the analysis of identifiability problems were completed as early as in the 90s. Jacquez and Perry [274] applied Fortran routines to several biological models. Some drawbacks were present though, because models should be coded in the Fortran programming language and the analysis was performed only at local values of the parameter space. If we search for a tool compatible with SBML models, a very useful software provided by Rodriguez-Fernandez and Banga [473] is `SensSB`, a toolbox for MATLAB that calculates both local and global sensitivity indexes and connects them with identifiability and optimal experimental design. *A priori* and *a posteriori* identifiability analysis has been also implemented in `PottersWheel` [249, 459], a MATLAB toolbox. Finally, `DAISY` [57] is to our knowledge the only software available for *a priori* global identifiability, unfortunately, not compatible with SBML.

As biochemical models are getting larger and larger, efficiency is an important aspect to consider for modelling tools. Not many tools allow parallel execution. `Grid Cellware` [134] allows to perform deterministic and stochastic simulations and model calibrations in grid environments. If our model contains thousands of reactions and species, `Hy3S` [494] has been proved to be an efficient tool to study such models. Simulations can be treated in a hybrid manner distinguishing the deterministic from the stochastic part.

Finally I want to emphasise that the current section is intended to introduce the results presented in this work. `ByoDyn` is a computational tool within the context explained here. Some of its features are unique among all other tools, other are scattered along different tools, others are shared. The motivation for the construction of such a tool was clear: the power to develop the features that meet our needs. Currently the tool offers a rich bunch of diverse types of analyses hard to find in a single tool. `ByoDyn` is explained thoroughly in Section 3.

2

Objectives

2.1 General Objective

The objectives of the current PhD thesis can be framed in a quantitative approach to biology. Biological systems, as any other existing reality, can be treated quantitatively. Historically the field of molecular biological suffer from difficulties to treat their matter of research in an accurate quantitative way. Thanks to the important advances of experimental techniques, the data acquired from biochemical systems has been on demand of computational treatment. The general objective of the work presented here is to develop and apply appropriate computational tools for the understanding and interpretation of biological systems. The choice and application appropriate methods to study biological systems at the different levels of organization of cellular systems has been a constant *motto* all along the work uncovered here. Secondary, at the time this work was conceived, the opening of an entire new line of research within the group was fostered.

2.2 Specific Objectives

The specific objectives of this PhD thesis are:

2.2. SPECIFIC OBJECTIVES

1. Study the topological features of cellular networks, both at the interaction and expression level.

2. Own a computational framework for the quantitative analysis of the dynamical features of mathematical models of biochemical interest.

3. Apply our computational framework, ByoDyn, to biologically relevant models.

Along the following section the specific objectives will be set forth in detail.

El primer escalón
siempre es el más difícil.
Los abrazos rotos, 2009

3 Results

3.1 General Overview

Within the group of Computational Biology and Biophysics Lab, I started my PhD thesis project in January 2004. At the time, the main focus of research of the group was in the microscopic aspects of biological systems, specifically, biochemical reactivity, protein folding and molecular interactions. Then, the group decided to open new fields of research on the macroscopic aspects of biochemical systems. At that point, the group decided to glue the experiences and knowledge on biochemistry from one side and on computer modelling from the other.

Following this path we started a collaboration with the Developmental Biology Group of Pompeu Fabra University to better understand the pattern formation arising on the otic placode during chick organogenesis. The system in hands was fairly large involving many genes and proteins regulating each other on different cells. Efficient model calibration on a multicellular system constituted a difficult handicap for the available tools at the current time. Facing this condition we decided to build our own code for the analytical purposes we had in mind. Within this framework `ByoDyn` was born. Observing the necessity of good (in the sense of universal) coding system for the models studied, we incorporated SBML [268] compatibility on `ByoDyn`. The compatibility was implemented during a short stay at the Control and Dynamical System Laboratory in Caltech under the supervision of Mike Hucka. At this

point we became aware of the large and steadily growing variety of models available in SBML, specifically, from public repositories as BioModels Database [322]. Given the different relevant aspects of the models, different analytical approaches should be taken which brought us to develop different algorithms in ByoDyn: from deterministic to stochastic simulators, sensitivity analysis, model calibration, parallel processing, optimal experimental design and many more. Today's situation of ByoDyn can be stated as a largely complete tool for the analysis of mathematical models of biochemical interest.

While ByoDyn reached a mature status we started the application of our tool to specific systems. The model of lateral inhibition at the otic placode represented a good ground for the application of the algorithms included in ByoDyn. One of the major problems of the system under study is the experimental difficulties to obtain precise and frequent samples along the time scale of the experiment. On this direction, our motivation was to generate a computational protocol to determine which measurements (which species and at which time) resulted more appropriate for the model calibration problem. Finally, due to the computational cost of the complete protocol, which includes rounds of model calibration, some key algorithms of ByoDyn were parallelised. A short stay for six weeks, funded by the HPC-Europa granting frame (http://www.hpc-europa.eu/) under the supervision of Gianni de Fabritiis at University College London was essential for the completion of this goal.

Finally we were interested on another critical aspect of the biochemical networks: topology. Genes, proteins and metabolites interact together in the cellular context setting up a network of connections. Generally speaking, the typical size of these networks is in the order of thousands for both elements and interactions. Obviously then, statistical methods are required to determine unequivocally the properties of such networks. With this aim I spend two months at Zoltán Oltvai's laboratory at University of Pittsburgh where we applied appropriate statistical methods to a large variety of experimental data from cellular networks. Either retrieved from public repositories or generated at Oltvai's lab, the data selected was as most up to date and accurate as possible. The data analysed reflected the genome at different conditions, the proteome, the transcriptome and the metabolome; both activity and interaction aspects

of the networks were studied. The results of the work revealed that the established models for network generation fail to explain some of the main properties of the cellular networks.

More specifically we summarise and present in the next section the articles that constitute the results of the thesis.

3.2 Papers

3.2.1 Paper 1. ByoDyn: A Tool for the Computational Analysis of Biochemical Kinetic Models

In this paper we present a new computational tool for the analysis of mathematical models of biochemical interest. ByoDyn reads models in SBML format [268], which is the *de facto* standard format [144] for computational models of biochemical systems. We benefit from that fact by accessing a broad variety of models, stored in public repositories like BioModels Database [322] and JWS Online Model Database [422]. Our platform is capable of a full range of analysis protocols: deterministic and stochastic simulations, model calibration, sensitivity analysis, identifiability analysis, optimal experimental design and many others. ByoDyn is able to take profit of different architectures as it runs in single PCs, clusters and supercomputers. Some of the model calibration algorithms have been parallelised and ByoDyn has been deployed into the QosCosGrid [312] platform which offers supercomputer behaviour on clusters. Finally in order to facilitate usability, a public web-based graphical interface called ByoDynWeb has been developed as an alternative access to the program.

BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–2

ByoDyn: A Tool for the Computational Analysis of Biochemical Kinetic Models

Adrián López García de Lomana, Àlex Gómez-Garrido, Miguel Hernández-Sánchez, Pau Rué, Jordi Villà-Freixa*†

CBBL, Research Unit on Biomedical Informatics, IMIM/Universitat Pompeu Fabra, C/Doctor Aiguader, 88, 08003 Barcelona, Catalunya, Spain

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: In ByoDyn, SBML models can be imported, visualised, simulated and analysed using the methods provided. Both deterministic and stochastic simulations can be performed, and systems can be analysed by means of parameter estimation, sensitivity and identifiability analysis, and optimal experimental design. The web interface provides a personalised workspace and transparent links to calculations in different computational environments.

Availability: http://byodyn.imim.es provides links to both the web GUI server and the GPL source code.

Contact: jordi.villa@upf.edu

1 INTRODUCTION

The development of standards for systems biology (SB) has led to an explosion of available models and applications for this field [Klipp et al., 2007]. The field is now mature enough to make use of advanced algorithmics beyond the basic simulation schemas provided by deterministic and stochastic frameworks. Thus, a researcher may be interested in estimating parameters for a model from a set of experimental data, analysing the identifiability of such parameters, or even proposing new experiments through different flavours of optimal experimental design techniques [Kreutz and Timmer, 2009, Banga and Balsa-Canto, 2008, Chen et al., 2009]. Equally relevant is the need to provide tools to easily integrate software from different disciplines in a transparent way for the user, and here web services are becoming a standard way to intercommunicate tools in, sometimes, distant servers.

2 FEATURES

ByoDyn can import SBML models through its binding to libSBML [Bornstein et al., 2008]. The program numerically solves the dynamics of species for models grounded in different deterministic frameworks (ODEs, DDEs, DAEs), including events, rules and user defined functions. The appropriate solver is automatically selected from Python bindings to SciPy, Octave, XPP or OpenModelica (for a complete list of web sites for these and other tools used within ByoDyn, please, refer to the supplementary material). Stochastic simulations are run using the direct method implementation of Gillespie's stochastic simulation algorithm (SSA) or different variants of τ–leaping methods [Gillespie, 2007, Rué et al., 2009].

Different parameter estimation tools are available, including local (gradient-based, using methods from the Netlib PORT library), global (genetic algorithms, GA) or hybrid (different combinations of the former) optimization methods. Sensitivity analysis [Stelling et al., 2004], identifiability analysis and optimal experimental design (OED) based on the Fisher information matrix [Banga and Balsa-Canto, 2008] can be performed with the program. The Python-based object oriented structure of the code permits a modular implementation of the different tasks and paves the way for easy further implementation following the open source philosophy. Graphics are obtained through the link to gnuplot, matplotlib or by direct PostScript/PNG outputs.

2.1 Web Interface and distributed computing

Focussing on the biomedical field, the ultimate goal of SB should be to provide links between basic and clinical research and to this end the integration of its methods into biomedical informatics workflows is going to become critical in the years to come. Thus, it is important to provide easy to implement links between knowledge management portals and SB tools. It is also necessary to provide easy to use graphical interfaces, aiming not just at expert SB users but also at experimentalists willing to test their hypothesis in an integrated environment. To this end, access to ByoDyn is both command-line based and web based.

The graphical web based interface includes support for systems biology graphical notation (SBGN) visualization and editting of models, while providing an easy link to popular SBML model repositories and a personal workspace with a UNIX-like sharing system for input files, results and models, helping creating collaborative environments (due to computational restrictions, guest access is provided with some limitations on the accessible features). The web GUI has been developed on top of a *LAMP* system and web services are being used to import ByoDyn from wider scope portals, e.g., BioBridge (http://www.biobridge.eu). Figure 1 shows how the diverse program features can be reached from the web GUI. Documentation, model repository links and examples are also provided through the web GUI.

*to whom correspondence should be addressed

†ALGL and AGG contributed equally to the work, MHS developed the web interface and PR developed the stochastic simulations code.

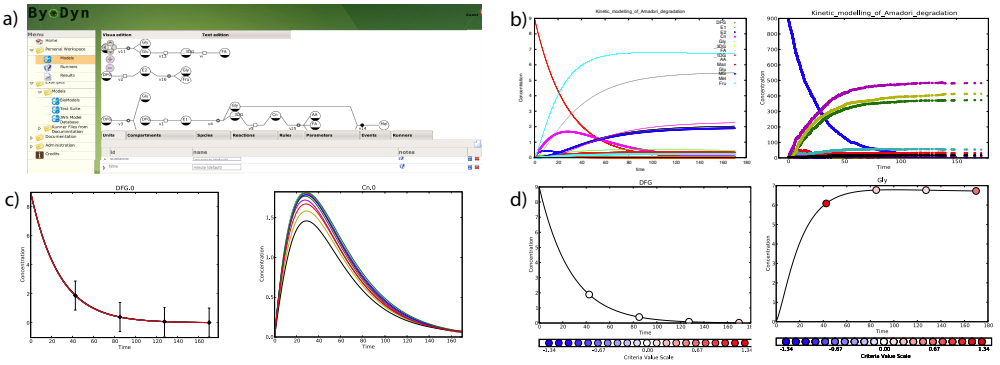

Fig. 1. Snapshots of the ByoDyn web GUI. Taking as an example the model in Martins and Van Boekel [2003], the figure shows: a) visualisation/editting of SBML files following the SBGN standards; b) deterministic and stochastic integration; c) reconstruction of the trajectories obtained after parameter estimation (local method starting from 10 initial parameter vectors randomly sampled within two orders of magnitude centered on the model nominal values). As a proof of concept, the calculation used 4 computer generated points as pseudo-experiments. Note the difference between the behaviour of the targetted nodes (left) and the non-targetted nodes (right); and d) results of the local OED protocol using the modified E criterion[Banga and Balsa-Canto, 2008], showing the information associated with each of the data points used in the parameter estimation protocol.

An important aspect of the program is its capability to run in different computational environments, depending on the user permissions, which is done in a transparent way for the user (local cluster, external high-performance computing resources). This includes the possibility of running parallel computations through the standard message passing interface (MPI) in ScientificPython (including its deployment on the QosCosGrid environment (`http://www.qoscosgrid.eu`), while providing basic computing capabilities for guest users.

The supplementary material for this paper includes a complete user's reference, describing all ByoDyn features in more detail, and a quick start guide.

2.2 Future prospects

Although there exist tools able to translate CellML into SBML files [Schilstra et al., 2006] and *vice versa*, the two standards do not share 100% of their features. Thus, the upcoming versions of our program are scheduled to provide support for CellML. There will be also support for steady state analysis, although current research in the lab is focussed on the improvement of identifiability of parameters by means of global analysis of the parameter landscape and multiscale simulations from coupled deterministic-stochastic approaches.

FUNDING

This work was supported by the Spanish Ministerio de Ciencia e Innovación [grant number CTQ2008-00755/BQU]; La Caixa graduate program [to PR]; Generalitat de Catalunya [to ALGL]; and European Commission [grant numbers FP6-2005-LIFESCIHEALTH-7 037909, FP6-2005-IST-5 033883, FP7-2007-IST-223920].

ACKNOWLEDGEMENTS

We thank M. Johnston, M. de Cáceres and D. Sportouch for support in different stages of the development of this work, M. Hucka and B. Bornstein for SBML i/o support and S. Laurie for English revision.

REFERENCES

JR Banga and E Balsa-Canto. Parameter estimation and optimal experimental design. *Essays Biochem.*, 45:195–209, 2008.

BJ Bornstein, SM Keating, A Jouraku, and M Hucka. LibSBML: an API library for SBML. *Bioinformatics*, 24:880–881, 2008.

WW Chen, B Schoeberl, PJ Jasper, M Niepel, UB Nielsen, DA Lauffenburger, and P K Sorger. Input-output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data. *Mol Syst Biol*, 5:239, 2009.

DT Gillespie. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem*, 58: 35–55, 2007.

E Klipp, W Liebermeister, A Helbig, A Kowald, and J Schaber. Systems biology standards–the community speaks. *Nat Biotechnol*, 25:390–391, 2007.

C Kreutz and J Timmer. Systems biology: experimental design. *FEBS J.*, 276:923–942, 2009.

SI Martins and MA Van Boekel. Kinetic modelling of Amadori N-(1-deoxy-D-fructos-1-yl)-glycine degradation pathways. Part II–kinetic analysis. *Carbohydr. Res.*, 338: 1665–1678, 2003.

P Rué, J Villà-Freixa, and K Burrage. Tau leap fully explicit runge-kutta methods with extended stability regions for stiff biochemical kinetics. *J. Chem. Phys.*, submitted, 2009.

MJ Schilstra, L Li, J Matthews, A Finney, M Hucka, and N Le Novre. CellML2SBML: conversion of CellML into SBML. *Bioinformatics*, 22:1018–1020, 2006.

J Stelling, ED Gilles, and FJ Doyle III. Robustness properties of circadian clock architectures. *PNAS*, 101:13210–13215, 2004.

3.2.2 Paper 2. Optimal Experimental Design in the Modelling of Pattern Formation

In this work we apply optimal experimental design techniques [162, 475] to a multicellular model of pattern formation in developmental biology [377, 17]. Apart from exceptional results [55], kinetic parameters of biochemical systems are not easily measurable. Instead, if sufficient experimental data of key species is available, kinetic parameters' values can be inferred using model calibration techniques [39]. Generally, experimental data about the output of the system is scarce [514] and measured at time points decided by the experimentalists without any prior knowledge. We apply here a complete protocol for optimal experimental design which provides information about which would be the next measurement so that the model calibration problem becomes easier. The protocol is based on the local improvement of the objective function shape [475]. We have detected a main drawback along the iterations of the protocol: given the roughness of the objective function and therefore the abundance of many local minima, we risk at each minimisation procedure to end up in a different minimum. This problem validates the application of the complete protocol only at the vicinity of the parameter nominal values. Otherwise a global characterisation of the objective function is required [96, 173, 626].

López A, Gómez-Garrido A, Sportouch D, Villà-Freixa J. Optimal Experimental Design in the modelling of pattern formation. Lect Notes Comput Sci. 2008; 5101: 610-19.

3.2.3 Paper 3. Statistical Analysis of Global Connectivity and Activity Distributions in Cellular Networks

In this article we use appropriate mathematical methods to determine some of the main topological features of cellular networks. We can account for mainly three types of interaction networks in the cellular context: protein-protein interaction networks (PIN), transcriptional regulatory networks (TRN) and metabolic networks (MN). All of them have been said [279, 165, 280] to follow power-law distributions for the connectivity degree. Moreover, the activity of cellular networks have been declared to follow power-law distributions [168, 568, 10]. Elicited by other works [296, 543, 550, 602] which raised methodological problems on the analysis of the networks' degree distributions, we analysed in this paper a broad number of the most accurate and complete data sets on cellular networks using appropriate statistical methods [103]. The networks we analysed were PIN (retrieved from literature curated sources or high-throughput experiments [464]) and TRN [355] in yeast and the *E. coli* reactome [169]. We also studied gene expression data in *E. coli* at different growth conditions [54] and protein expression data [200]. All tested networks either differed significantly from the power-law distribution or alternative models were equally good explaining the empirical distributions. Thereafter we concluded that none of the studied networks can be accounted unequivocally for power-law on the connectivity degree distribution.

López A, Beg QK, De Fabritiis G, Villà-Freixa J. Statistical analysis of global connectivity and activity distributions in cellular networks. J Comput Biol. 2010; 17(7): 869-78

There is only one thing in the world worse than being talked about, and that is not being talked about.

Oscar Wilde

4 Discussion

4.1 Cellular Networks and Graphs

Since little before the 2000s decade, catalysed by some seminal works as [592, 43], there has been a prominent interest on the study of the properties of real networks. Apart from the extraordinary work of theoretical approaches in graph theory, the burst in the field has been a consequence of the availability of interaction data from real networks. For example, in sociology, relationships were established based on personal interviews which are time consuming and subjective. Nowadays, however, on-line social networks or e-mail communications provide with a source of data much more reliable and retrievable in extremely short time. In the biochemical arena, high-throughput techniques and text mining tools have uncovered an unprecedented large amount of biochemical interaction data: protein-protein physical interactions, gene regulation relationships and connections among metabolic reactions. However, the accuracy of high-throughput methods is far from the ideal [149], in fact, for yeast two hybrid (Y2H) experiments, the false positive rates are up to 64% and false negatives from 43% to 72% [317]. Other high-throughput methods employed, like tandem affinity purification (TAP) experiments have also very low reliability with 77% for false positives and 15%-50% for false negatives [317]. With this type of data, the derivation of accurate predictive theories about the behaviour of biochemical networks results utopian. Nevertheless, the use of more accurate data sets [464], derived from example from literature curation and the use of sta-

tistical methods to assess the confidence of the analyses [317] represents a secure path for future research.

Indeed, in the field of network biology several myths have extended [340]: (1) power-law degree distribution, (2) scale-freeness, (3) small-world-ness, (4) error tolerance simultaneous to attack vulnerability and (5) preferential attachment as the network growth mechanism. In order to have greater insights on the first hypothesis, i.e., the power-law degree distribution of cellular networks, we have performed the most appropriate statistical analysis to the state of the art biochemical data sets (See Section 3.2.3) finding no specific support for the power-law model. In the case the results of the analysis would lean to a specific distribution, additional remarks would be taken into consideration as small size data sets or inherent noise. Practices of leave-one-out cross validation would provide, in that case, with robust conclusive results.

From a biochemical perspective, node functional inference or prediction of global network behaviour results more appealing than the determination of the probability distribution of a graph feature. Then, whether or not biochemical networks display power-law degree distribution, other properties as community structure, largely linked to functional modules, can arise large interest from biochemistry scientists. Is in this direction already described or newly discovered graph measures should be tested in cellular networks. Intuitively, the importance of community structure in metabolic networks is considerable, in fact, metabolism is organised in cycles, cascades or pathways in general, which tend to be relatively distinguishable. In this framework, however, cross-talk among pathways is essential for a proper regulation of the cellular responses. Key players connecting several pathways identified conceptually as *date hubs* in [239] or more accurately as R3, R6 and R7 type nodes in [231, 233] should be identified and characterised to better understand, as a whole, the complete set of relationships[1].

Graph analysis tools have proved to be largely useful for biochemistry, specially for the analysis of *omics* data sets. Specifically, as it has been

[1]Metaphorically, certain nodes touch different network communities as motorways link different large cities or states, while other nodes touch only intra-community nodes, as secondary roads link nearby villages. The characterisation of those nodes that bind different communities can be a proxy for the global picture of the network.

exposed in Section 1.2.1, many algorithms have been developed to infer community structure in graphs. The uncovering of communities in graphs derived from biochemical interactions allows for further biological knowledge, as closely clustered biological entities can be assumed to be involved in the same biological process. Nevertheless, a recent work from Song and Singh [529] have shown that function prediction is better accomplish in the *S. cerevisiae* PIN by a guilt-by-association approach than by several leading edge clustering algorithms. Other works [291] emphasise the lack of relationship between strength of community structure and robustness of community structure: some graphs may display a largely significant community structure but small perturbations may vanish it completely. Consequently, a word of caution should be addressed when deriving biological attributes from graph features of biochemical networks.

Basic scientific disciplines like physics or chemistry has largely founded the knowledge in biochemistry and molecular biology, aiding in the understanding of, for example, the activity of proteins or the kinetics of enzymes. Graph theory too, has proved to be another helpful tool to take into account in biochemistry and many other further insights are expected.

4.2 Dynamical Analysis of Mathematical Models of Biochemical Interest

All scientific disciplines concerned about understanding nature combine together experimental approaches with rigorous analytical approaches. Which formal methods and how we use them is critical for the understanding of biological world. The work of Bashor *et al.* [47] could be one of the innumerable examples. The usefulness of the models we work with relies not on their ability to reproduce the observed behaviour but to predict novel results. Which mathematical formalism we choose is the first important matter we have to solve. There is not a universal rule to choose a formalism or another, being up to the choice of the scientist. The main idea to use the simplest model without missing the features we are studying. In fact, the mathematical framework we use is of crucial im-

portance and the results derived from different formalisms might bring us to different predictions and conclusions. A clear example would be the choice between deterministic or stochastic methods. Understanding cellular behaviour is the ultimate goal of systems biology and cellular behaviour is the result of stochastic molecular interactions [278]. Although historically continuous deterministic assumptions have been considered for modelling biochemical systems, efforts to determine sensitivity [115] or even to perform model calibration [418, 188] in stochastic systems have been conducted.

At any rate, independently of the method we use, i.e., whether we model our system using discrete or continuous approaches, deterministic or stochastic, we need to validate our model. At the validation stage, from a more philosophical perspective, we have to keep in mind the question: *What we have built the model for?* or *Which questions we wanted to answer?*. Within this context, the objective function will never be an adequate measure for goodness of the model. Instead, if for example, our goal was the determination of the parameter values from dynamical data, identifiability analysis provide us with adequate answers on model validation.

Model calibration, which is a central problem in systems biology and for which much resources have been allocated, requires from both computational and the experimental efforts. Methods for model calibration are generally computationally too expensive turning into rather slow options; on the other hand, acquisition of temporal quantitative data for biochemical species in cellular environments is highly complicated (except for some exceptions [439]). Data of low quality is the fist obviously burden for a successful model calibration. As a rule of thumb, the experimental data should be several times the number of parameters to estimate. A pervasive problem in model calibration of biological systems is the lack of adequate data [628]. Another important aspect is the noise of the measured data. While it might not always be possible, smoothing noisy data could help dramatically model calibration. Collaterally, working with smoothed resulting functions instead of noisy data sets [80] may help bypassing model integration, which is for sure the most computationally demanding task of model calibration: from 95% to nearly 100% of the resources is spent on model integration [576]. In this sense, efficient

algorithms for model integration and algebraic strategies are essential points of the model calibration problem. Model integration algorithms need to be fast and robust: fast because they are executed many times along the objective function minimisation and robust because different parameter values will be explored, yielding various dynamics [26].

Clearly, the choice of the model calibration method itself is a decisive issue. Local methods are fast and inconsistent for non-convex problems and global methods are more robust but they show slow convergence. Then, hybrid strategies (combination of both global and local phases) will generally be the best compromise. Typically, objective functions instead of soft surfaces (Figure 1.3), they are more likely to be majorly planar but with very deep and narrow basins of attraction (Figure 1.4). Such structures represent an extreme challenge for minimisation algorithms and the global minimum is missed oftenly. Importantly, a model calibration benchmark using time series data has been defined recently by Gennemark and Wedelin [198]. Given the computational cost of inverse problems, many of the available methods are commonly parallelised for efficiency reasons. However, worth noting is the argument from the authors that none of their tested problems within the field of systems biology required from supercomputing. Problems were solved at the range of hours in ordinary desktop computers. Not only model calibration but identification of the model structure has also been surveyed [313]. These works represent a major path for the development of more efficient model calibration algorithms for systems biology.

When model calibration fails, we start asking ourselves why. A relatively low value for the objective function only expresses the quality of the fit but does not provide any information about the quality of the estimated parameters [25]. Main sources of poor model calibration could be defined as [97]: the model, the data, the computational cost or any combination of all three. Taken the problem of computational cost apart, which can be alleviated using alternative minimisation algorithms, the combination of model and data can be studied with identifiability analysis. Identifiability analysis will eventually reveal some of the main problems of model calibration, e.g., flat regions on the objective function or convergence to local solutions[2]. As a matter of fact, the identifiability analysis should

[2]Not only identifiability analysis, but other techniques like global sensitivity indexes

be done prior to any model calibration. We should not need to ask ourselves why the model calibration failed, but rather controlling the sources of failure before any calculation is launched. The application of identifiability analysis is strongly recommended for models of biological interest, for which large emphasis is put into the biological interpretation of the estimated parameter values [26], being a must prior to model calibration.

For example Fomekong-Nanfack *et al.* worked on an initial model of 66 parameters. After fixing only four parameters the model calibration algorithms behave strikingly different, finding a relatively good result with much higher frequency. What does that mean? Which is the importance of those 4 parameters? And the importance of the other 62? How much importance show we put on those model parameters that show very wide distributions after model calibration? How much relevant should they be considered for the model? These types of questions, basically the relationship between $\boldsymbol{\theta}$ and $\mathcal{Y}$, bring us to the concept of sloppiness [236]. A large number of biological models studied, hold parameters such that changing their numerical values does not particularly affect its dynamic response. In those cases, modellers should focus on model predictions rather than model parameters [32, 236], specially for tentative or incomplete models. In sloppy models, parameters show a large variance while predictions remain tight; it is the dynamic response from ensembles of parameters which should be evaluated, not single points of the parameter space [75].

Almost all biochemical models for which identifiability analysis has been applied, showed several levels of unidentifiability, both *a priori* [618, 191, 190, 459, 36, 477] or *a posteriori* [459, 36]. Srinath and Gunawan [534] argument that for the majority of models using power-law formalisms, around 80% of the model parameters are *a priori* unidentifiable and 50-60% *a posteriori* unidentifiable. Worse, taking into account the typical noise of such biological measurements, only around 25% of the model parameters could be identified. Worth noting is that *a priori* unidentifiable parameters will not be recovered with better data sets [452]. In fact it is a central matter in biological modelling, not a merely peripheral issue. A positive point of view is that determination of a subset of parameters (the stiff ones) is sufficient to recover a consistent dynamic behaviour

[300] have been used to improve and understand model calibration results.

[94].

An interesting link can be drawn from parameter identifiability and model reduction. When it is detected a lack of parameter identifiability, then it is a good time for model reduction [510]. A general trend in biochemical modelling is to build more and more complex models, with an increasing number of model parameters. However that does not necessarily imply better models, models capable to predict more trustfully. Specially if our models are data-driven, we should rely on the models that hold a succinct number of parameters such that different model hypothesis can be discriminated with the data in question.

Another important aspect which is not always sufficiently emphasised is that sloppy models do not necessarily correspond directly to biological properties. Identifiability is a mathematical notion, raising from the model. Whether if the biological process we are modelling is robust or not, is a different matter [185, 586]. It should be taken into consideration that most parameters of biochemical models are, with few exceptions, phenomenological.

In such a complicate context of parameter unidentifiability OED will not solve the problem but at least could guide us in the determination of the most useful experiment. At least, OED will tell us which measurements are necessary to fix the stiff parameters of the system [3], reducing human and economic costs [35]. Most of the literature presents works on OED based on the FIM. The FIM is evaluated at a single point of the parameter space and more robust methods would be more suitable for the type of problems arising in biochemical modelling, which show non-convex objective functions. Moreover, models showing disparate identifiable parameters (sloppy) tend to generate singular FIMs, making FIM-based criteria useful only for identifiable models. Additionally, if we are interested on applying protocols of OED, the parameter nominal values are generally unknown and ideal OED techniques should deal with that. For those reasons, some efforts have been developed in the direction of global OED. Hengl *et al.* [249] have studied the parameter relations on

[3] An ultimate extension of OED for model calibration would be OED for model selection: defining which $\mathcal{Y}$ is best for discriminating among models of different structure.

a distribution of model calibration fits. Another example is the work of Feng *et al.* [175] who generated indexes of OED based on the limits of the distribution of parameters derived from model calibration. From my personal perspective, although computationally costly, OED indexes based on information theoretic measures, like Shannon entropy, would yield largely trustful results.

As final point the software tool `ByoDyn` has been presented. `ByoDyn` is fairly complete tool which attempts to deal with most of the problems discussed above. It is based on computational standards like SBML or MPI and intended to be easily accessible and free for the scientific community: none of its dependencies is under commercial licence.

5

Conclusions

The conclusions of the PhD thesis are the following:

1. Cellular networks, both at the level of interaction and activity display heavy-tailed distributions. However, using adequate data sets of interaction and activity networks in yeast and *E. coli* and appropriate statistical methods, the resulting distributions could not be assigned unequivocally to any of the suggested distributions, included the power-law.

2. A new computational tool `ByoDyn` has been developed from scratch. `ByoDyn` is fairly complete tool which attempts to deal with most of the problems encountered on the study of dynamical models of biochemical interest. Built from computational standards, its idiosyncrasy is to be easily accessible and free for the scientific community. Understanding quantitatively the processes occurring in a cell requires from the computational tools found in `ByoDyn`.

3. Despite of their computational cost, prevalent methods for the dynamical analysis of mathematical models of biochemical interest, e.g. sensitivity analysis, model calibration, identifiability analysis and optimal experimental design, need for global strategies for a consistent description of the dealing problems. Lack of convergence is generally suffered from local approaches applied to dynamical aspects of biochemical modelling.

Bibliography

[1] Adamcsek, B., Palla, G., Farkas, I. J., Derenyi, I. and Vicsek, T. *CFinder: locating cliques and overlapping modules in biological networks.* Bioinformatics **22**:1021–1023 (2006).

[2] Adamic, L. A., Huberman, B. A., Barabási, A., Albert, R., Jeong, H. and Bianconi, G. *Power-law distribution of the world wide web.* Science **287**:2115 (2000).

[3] Agarwal, S., Deane, C. M., Porter, M. A. and Jones, N. S. *Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks.* PLoS Comput. Biol. **6**:e1000817 (2010).

[4] Alava, M. J. and Dorogovtsev, S. N. *Complex networks created by aggregation.* Phys. Rev. E **71**:036107 (2005).

[5] Alber, M., Hentschel, H., Kazmierczak, B. and Newman, S. *Existence of solutions to a new model of biological pattern formation.* J. Math. Anal. Appl. **308**:175–194 (2004).

[6] Albert, R. *Scale-free networks in cell biology.* Journal of Cell Science **118**:4947–4957 (2005).

[7] Albert, R. and Barabási, A.-L. *Topology of Evolving Networks: Local Events and Universality.* Physical Review Letters **85**:5234–5237 (2000).

[8] Albert, R., Jeong, H. and Barabási, A.-L. *Diameter of the World-Wide Web.* Nature **401**:130 (1999).

[9] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. *Molecular Biology of the Cell, Fourth Edition.* Garland (2002).

[10] Almaas, E., Kovács, B., Oltvai, Z. N. and Barabási, A.-L. *Global organization of metabolic fluxes in the bacterium* Escherichia coli. Nature **427**:839–843 (2004).

[11] Alvarez-Hamelin, I., Dall'Asta, L., Barrat, A. and Vespignani, A. *k-core decomposition: a tool for the visualization of large scale networks.* CANADA page 41 (2006).

[12] Alves, R., Antunes, F. and Salvador, A. *Tools for kinetic modeling of biochemical networks.* Nat. Biotechnol. **24**:667–672 (2006).

[13] Amaral, L. A. N., Scala, A., Barthélémy, M. and Stanley, H. E. *Classes of small world networks.* PNAS **97**:11149–11152 (2000).

[14] Ander, M., Beltrao, P., Di Ventura, B., Ferkinghoff-Borg, J., Foglierini, M., Kaplan, A., Lemerle, C., Tomas-Oliveira, I. and Serrano, L. *SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks.* Syst. Biol. **1**:129–138 (2004).

[15] Andrews, S. S., Addy, N. J., Brent, R. and Arkin, A. P. *Detailed simulations of cell biology with Smoldyn 2.1.* PLoS Comput. Biol. **6**:e1000705 (2010).

[16] Angeli, D., Ferrell, J. E. and Sontag, E. D. *Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems.* Proc. Natl. Acad. Sci. U.S.A. **101**:1822–1827 (2004).

[17] Appel, B., Givan, L. A. and Eisen, J. S. *Delta-Notch signaling and lateral inhibition in zebrafish spinal cord development.* BMC Dev. Biol. **1**:13 (2001).

[18] Apri, M., Molenaar, J., de Gee, M. and van Voorn, G. *Efficient estimation of the robustness region of biological models with oscillatory behavior.* PLoS ONE **5**:e9865 (2010).

[19] Arenas, A., Danon, L., Díaz-Guilera, A., Gleiser, P. and Guimerà, R. *Community analysis in social networks.* The European Physical Journal B **38**:373–380 (2004).

[20] Arenas, A., Díaz-Guilera, A. and Pérez-Vicente, C. J. *Synchronization reveals topological scales in complex networks.* Phys. Rev. Lett. **96**:114102 (2006).

[21] Arenas, A., Fernández, A. and S., G. *Analysis of the structure of complex networks at different resolution levels.* New J. Phys. **10**:053039 (2008).

[22] Arita, M. *The metabolic world of* Escherichia coli *is not small.* PNAS **101**:1542–1547 (2004).

[23] Arkin, A., Ross, J. and McAdams, H. H. *Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected* Escherichia coli *cells.* Genetics **149**:1633–1648 (1998).

[24] Arkin, A., Shen, P. and Ross, J. *A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements.* Science **277**:1275–1279 (1997).

[25] Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J. A. and Blom, J. G. *Systems biology: parameter estimation for biochemical models.* FEBS J. **276**:886–902 (2009).

[26] Ashyraliyev, M., Jaeger, J. and Blom, J. G. *Parameter estimation and determinability analysis applied to* Drosophila *gap gene circuits.* BMC Syst. Biol. **2**:83 (2008).

[27] Ashyraliyev, M., Siggens, K., Janssens, H., Blom, J., Akam, M. and Jaeger, J. *Gene circuit analysis of the terminal gap gene huckebein.* PLoS Comput. Biol. **5**:e1000548 (2009).

[28] Asprey, S. P. and Macchietto, S. *Designing robust optimal dynamic experiments.* Journal of Process Control **12**:545–556 (2002).

[29] Atkinson, A. and Donev, A. *Optimum Experimental Designs.* Clarendon Press (1992).

[30] Audoly, S., Bellu, G., D'Angio, L., Saccomani, M. P. and Cobelli, C. *Global identifiability of nonlinear models of biological systems.* IEEE Trans. Biomed. Eng. **48**:55–65 (2001).

[31] Bachar, M. and Dorfmayr, A. *HIV treatment models with time delay.* C. R. Biol. **327**:983–994 (2004).

[32] Bailey, J. E. *Complex biology with no parameters.* Nat. Biotechnol. **19**:503–504 (2001).

[33] Baker, C., Bocharov, G., Paul, C. and Rihan, F. *Computational modelling with functional differential equations: Identification, selection, and sensitivity.* Applied Numerical Mathematics **53**:107–129 (2005).

[34] Baker, S. M., Schallau, K. and Junker, B. H. *Comparison of different algorithms for simultaneous estimation of multiple parameters in kinetic metabolic models.* J. Integr. Bioinform. **7**(3):133 (2010).

[35] Balsa-Canto, E., Alonso, A. A. and Banga, J. R. *Computational procedures for optimal experimental design in biological systems.* IET Syst. Biol. **2**:163–172 (2008).

[36] Balsa-Canto, E., Alonso, A. A. and Banga, J. R. *An iterative identification procedure for dynamic modeling of biochemical networks.* BMC Syst. Biol. **4**:11 (2010).

[37] Balsa-Canto, E., Peifer, M., Banga, J. R., Timmer, J. and Fleck, C. *Hybrid optimization method with general switching strategy for parameter estimation.* BMC Syst. Biol. **2**:26 (2008).

[38] Bandara, S., Schloder, J. P., Eils, R., Bock, H. G. and Meyer, T. *Optimal experimental design for parameter estimation of a cell signaling model.* PLoS Comput. Biol. **5**:e1000558 (2009).

[39] Banga, J. R. *Optimization in computational systems biology.* BMC Syst. Biol. **2**:47 (2008).

[40] Banga, J. R., Balsa-Canto, E., Moles, C. G. and Alonso, A. A. *Dynamic optimization of bioprocesses: efficient and robust numerical strategies.* J. Biotechnol. **117**:407–419 (2005).

[41] Banga, J. R., Versyck and Van Impe, J. F. *Computation of Optimal Identification Experiments for Nonlinear Dynamic Process Models: a Stochastic Global Optimization Approach.* Ind. Eng. Chem. Res. **41**:2425–2430 (2002).

[42] Barabási, A.-L. *Network biology: understanding the cell's functional organization.* Nat. Rev. Genet. **5(2)**:101–13 (2004).

[43] Barabási, A.-L. and Albert, R. *Emergence of Scalling in Random Networks.* Science **286**:509–512 (1999).

[44] Barkai, N. and Leibler, S. *Robustness in simple biochemical networks.* Nature **387**:913–917 (1997).

[45] Barrat, A. and Weigt, M. *On the properties of small-world network models.* The European Physical Journal B-Condensed Matter **13(3)**:547–560 (2000).

[46] Barrio, M., Burrage, K., Leier, A. and Tian, T. *Oscillatory regulation of Hes1: Discrete stochastic delay modelling and simulation.* PLoS Comput. Biol. **2**:e117 (2006).

[47] Bashor, C. J., Helman, N. C., Yan, S. and Lim, W. A. *Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics.* Science **319**:1539–1543 (2008).

[48] Batada, N. N., Hurst, L. D. and Tyers, M. *Evolutionary and physiological importance of hub proteins.* PLoS Comput. Biol. **2**:e88 (2006).

[49] Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hurst, L. D. and Tyers, M. *Stratus not altocumulus: a new view of the yeast protein interaction network.* PLoS Biol. **4**:e317 (2006).

[50] Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hurst, L. D. and Tyers, M. *Still stratus not altocumulus: further evidence against the date/party hub distinction.* PLoS Biol. **5**:e154 (2007).

[51] Bazil, J. N., Buzzard, G. T. and Rundell, A. E. *Modeling mitochondrial bioenergetics with integrated volume dynamics.* PLoS Comput. Biol. **6**:e1000632 (2010).

[52] Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. . and Herrgard, M. J. *Quantitative prediction of cellular metabolism with constraint-based models: the* `COBRA Toolbox`. Nat. Protoc. **2**:727–738 (2007).

[53] Becskei, A. and Serrano, L. *Engineering stability in gene networks by autoregulation.* Nature **405**:590–593 (2000).

[54] Beg, Q. K., Vázquez, A., Ernst, J., de Menezes, B.-J. Z., M. A., Barabási, A.-L. and Oltvai, Z. N. *Intracellular crowding defines the mode and sequence of substrate uptake by* richia coli *and constrains its metabolic activity.* PNAS **104**:12663–12668 (2007).

[55] Belle, A., Tanay, A., Bitincka, L., Shamir, R. and O'Shea, E. K. *Quantification of protein half-lives in the budding yeast proteome.* PNAS **103**:13004–13009 (2006).

[56] Bellen, A. and Zennaro, M. *Numerical Methods for Delay Differential Equations.* Clarendon Press (2003).

[57] Bellu, G., Saccomani, M. P., Audoly, S. and D'Angio, L. *DAISY: a new software tool to test global identifiability of biological and physiological systems.* Computer Methods and Programs in Biomedicine **88**:52–61 (2007).

[58] Bender, E. A. and Canfield, E. R. *The asymptotic number of labeled graphs with given degree sequences.* Journal of Combinatorial Theory **Series A 24**:296–307 (1978).

[59] Bennett, D. and Gourley, S. *Periodic oscillations in a model of the glucoseinsulin interaction with delay and periodic forcing* . Dynamical Systems **19**(2):109–125 (2004).

[60] Bentele, M., Lavrik, I., Ulrich, M., Stosser, S., Heermann, D. W., Kalthoff, H., Krammer, P. H. and Eils, R. *Mathematical modeling reveals threshold mechanism in CD95-induced apoptosis.* J. Cell Biol. **166**:839–851 (2004).

[61] Berg, J., Lassig, M. and Wagner, A. *Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications.* BMC Evol. Biol. **4**:51 (2004).

[62] Bersini, H., Lenaerts, T. and Santos, F. C. *Growing biological networks: beyond the gene-duplication model.* J. Theor. Biol. **241**:488–505 (2006).

[63] Bertin, N., Simonis, N., Dupuy, D., Cusick, M. E., Han, J. D., Fraser, H. B., Roth, F. P. and Vidal, M. *Confirmation of organized modularity in the yeast interactome.* PLoS Biol. **5**:e153 (2007).

[64] Bianconi, G., Pin, P. and Marsili, M. *Assessing the relevance of node features for network structure.* PNAS **106**:11433–11438 (2009).

[65] Bilke, S. and Peterson, C. *Topological properties of citation and metabolic networks.* Phys. Rev. E **64**:036106 (2001).

[66] Bishop, C. *Neural Networks for Pattern Recognition.* Oxford University Press (1995).

[67] Bitner, A., Hołyst, R. and Fiałkowski, M. *From complex structures to complex processes: percolation theory applied to the formation of a city.* Phys. Rev. E **80**:037102 (2009).

[68] Blüthgen, N. and Herzel, H. *How robust are switches in intracellular signaling cascades?* J. Theor. Biol. **225**:293–300 (2003).

[69] Bocharov, G. and Rihan, F. *Numerical modelling in biosciences using delay differential equations.* Journal of Computational and Applied Mathematics **125**:183–199 (2000).

[70] Bock, H. G. *Recent advances in parameter identification techniques for ordinary differential equations.* Progress in Scientific Computing (Birkhäuser/Boston) **2**:95–121 (1983).

[71] Boettcher, S. and Percus, A. G. *Extremal optimization for graph partitioning.* Phys. Rev. E **64**:026114 (2001).

[72] Bollobás, B. *Random Graphs.* Academic Press, London (1985).

[73] Bongard, J. and Lipson, H. *Automated reverse engineering of nonlinear dynamical systems.* PNAS **104**:9943–9948 (2007).

[74] Brandes, U., Gaertler, M. and Wagner, D. *Experiments on graph clustering algorithms.* In *In 11th Europ. Symp. Algorithms* pages 568–579. Springer-Verlag (2003).

[75] Brown, K. S., Hill, C. C., Calero, G. A., Myers, C. R., Lee, K. H., Sethna, J. P. and Cerione, R. A. *The statistical mechanisms of complex signaling networks: nerve growth factor signaling.* Phys. Biol. **1**:184–195 (2004).

[76] Brown, K. S. and Sethna, J. P. *Statistical mechanical approaches to models with many poorly known parameters.* Phys. Rev. E **68**:021904 (2003).

[77] Brown, M., He, F. and Yeung, L. F. *Robust measurement selection for biochemical pathway experimental design.* Int. J. Bioinform. Res. Appl. **4**:400–416 (2008).

[78] Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G. and Chen, R. *Topological structure analysis of the protein-protein interaction network in budding yeast.* Nucleic Acids Res. **31**:2443–2450 (2003).

[79] Caldarelli, G., Capocci, A., De Los Rios, P. and Muñoz, M. A. *Scale-free networks from varying vertex intrinsic fitness.* Phys. Rev. Lett. **89**:258702 (2002).

[80] Cao, J. and Zhao, H. *Estimating dynamic models for gene regulation networks.* Bioinformatics **24**:1619–1624 (2008).

[81] Capocci, A., Servedio, V. D., Caldarelli, G. and Colaiori, F. *Communities Detection in Large Networks.* Lecture Notes in Computer Science **3243**:181–187 (2004).

[82] Cascante, M., Boros, L. G., Comin-Anduix, B., de Atauri, P., Centelles, J. J. and Lee, P. W. *Metabolic control analysis in drug discovery and disease.* Nat. Biotechnol. **20**:243–249 (2002).

[83] Casey, F. P., Baird, D., Feng, Q., Gutenkunst, R. N., Waterfall, J. J., Myers, C. R., Brown, K. S., Cerione, R. A. and Sethna, J. P. *Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model.* IET Syst. Biol. **1**:190–202 (2007).

[84] Casey, R., de Jong, H. and Gouze, J. L. *Piecewise-linear models of genetic regulatory networks: equilibria and their stability.* J. Math. Biol. **52**:27–56 (2006).

[85] Caulfield, E. and Hellander, A. *`CellMC`–a multiplatform model compiler for the Cell Broadband Engine and x86.* Bioinformatics **26**:426–428 (2010).

[86] Chaouiya, C. *Petri net modelling of biological networks.* Brief. Bioinformatics **8**:210–219 (2007).

[87] Chappell, M. J., Evans, N. D., Godfrey, K. R. and Chapman, M. J. *Structural Indetifiability of Controlled State Space Systems: A Quasi-Automated Methodology for Generating Indetifiabiable Reparameterisations of Unidentifiabiable Systems.* IEE Colloquium on Symbolic Computation for Control **8**:8/1–8/12 (1999).

[88] Chaves, M., Sontag, E. D. and Albert, R. *Methods of robustness analysis for Boolean models of gene control networks.* Syst. Biol. **153**:154–167 (2006).

[89] Chen, B. S., Wang, Y. C., Wu, W. S. and Li, W. H. *A new measure of the robustness of biochemical networks.* Bioinformatics **21**:2698–2705 (2005).

[90] Chen, J. and Yuan, B. *Detecting functional modules in the yeast protein-protein interaction network.* Bioinformatics **22**:2283–2290 (2006).

[91] Chen, K. C., Calzone, L., Csikasz-Nagy, A., Cross, F. R., Novák, B. and Tyson, J. J. *Integrative analysis of cell cycle control in budding yeast.* Mol. Biol. Cell **15**:3841–3862 (2004).

[92] Chen, M. and Hofestadt, R. *Quantitative Petri net model of gene regulated metabolic networks in the cell.* In Silico Biol. **3**:347–365 (2003).

[93] Chen, T., Filkov, V. and Skiena, S. S. *Identifying gene regulatory networks from experimental data.* In *RECOMB '99* pages 94–103. ACM Press (1999).

[94] Chen, W. W., Schoeberl, B., Jasper, P. J., Niepel, M., Nielsen, U. B., Lauffenburger, D. A. and Sorger, P. K. *Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data.* Mol. Syst. Biol. **5**:239 (2009).

[95] Chen, Y., Lawless, C., Gillespie, C. S., Wu, J., Boys, R. J. and Wilkinson, D. J. *`CaliBayes` and `BASIS`: integrated tools for the calibration, simulation and storage of biological simulation models.* Brief. Bioinformatics **11**:278–289 (2010).

[96] Cho, K.-H., Shin, S.-Y., Kolch, W. and Wolkenhauer, O. *Experimental Design in Systems Biology Based on Parameter Sensitivity Analysis Using a Monte Carlo Method: A Case Study for the TNFα-Mediated NF-κB Signal Transduction Pathway.* Simulation **79**:726–739 (2003).

[97] Chou, I. C. and Voit, E. O. *Recent developments in parameter estimation and structure identification of biochemical and genomic systems.* Math Biosci **219**:57–83 (2009).

[98] Chu, K., Deng, Y. and Reinitz, J. *Parallel simulated annealing by mixing of states.* J. Comput. Phys. **148**:646–662 (1999).

[99] Chu, Y., Jayaraman, A. and Hahn, J. *Parameter sensitivity analysis of IL-6 signalling pathways.* IET Syst. Biol. **1**:342–352 (2007).

[100] Chung, F., Lu, L., Dewey, T. G. and Galas, D. J. *Duplication models for biological networks.* J. Comput. Biol. **10**:677–687 (2003).

[101] Clauset, A., Moore, C. and Newman, M. E. J. *Hierarchical structure and the prediction of missing links in networks.* Nature **453**:98–101 (2008).

[102] Clauset, A., Newman, M. E. and Moore, C. *Finding community structure in very large networks.* Phys. Rev. E **70**:066111 (2004).

[103] Clauset, A., Shalizi, C. R. and Newman, M. E. J. *Power-law distributions in empirical data.* SIAM Review **51**:661 (2009).

[104] Cohen, S. D. and Hindmarsh, A. C. *Cvode, a stiff/nonstiff ode solver in C.* Computers in Physics **10**(2):138–143 (1996).

[105] Cokus, S., Rose, S., Haynor, D., Grønbech-Jensen, N. and Pellegrini, M. *Modelling the network of cell cycle transcription factors in the yeast Saccharomyces cerevisiae.* BMC Bioinformatics **7**:381 (2006).

[106] Colizza, V., Flammini, A., Serrano, M. A. and Vespignani, A. *Detecting rich-club ordering in complex networks.* Nature Physics **2**:110–115 (2006).

[107] Collier, J. R., Monk, N. A., Maini, P. K. and Lewis, J. H. *Pattern formation by lateral inhibition with feedback: a mathematical model of delta-notch intercellular signalling.* J. Theor. Biol. **183(4)**:429–446 (1996).

[108] Comin-Anduix, B., Boren, J., Martinez, S., Moro, C., Centelles, J. J., Trebukhina, R., Petushok, N., Lee, W. N., Boros, L. G. and Cascante, M. *The effect of thiamine supplementation on tumour proliferation. A metabolic control analysis study.* Eur. J. Biochem. **268**:4177–4182 (2001).

[109] Cooling, M. T., Rouilly, V., Misirli, G., Lawson, J., Yu, T., Hallinan, J. and Wipat, A. *Standard virtual biological parts: a repository of modular modeling components for synthetic biology.* Bioinformatics **26**:925–931 (2010).

[110] Coutinho, R., Fernandez, B., Lima, R. and Meyroneinc, A. *Discrete time piecewise affine models of genetic regulatory networks.* J. Math. Biol. **52**:524–570 (2006).

[111] Cukier, R., Fortuin, C., Shuler, K., Petschek, A. and Schaibly, J. *Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. Part I: theory.* Journal of Chemical Physics **59**(8):3873–3878 (1973).

[112] Culshaw, R. V. and Ruan, S. *A delay-differential equation model of HIV infection of CD4(+) T-cells.* Math. Biosci. **165**:27–39 (2000).

[113] Culshaw, R. V., Ruan, S. and Spiteri, R. J. *Optimal HIV treatment by maximising immune response.* J. Math. Biol. **48**:545–562 (2004).

[114] Cvijović, D. and Klinowski, J. *Taboo Search: An Approach to the Multiple Minima Problem.* Science **267**:664–666 (1995).

[115] Dacol, D. K. and Rabitz, H. *Sensitivity analysis of stochastic kinetic models.* J. Math. Phys. **25**:27162727 (1984).

[116] Dada, J. O., Spasić, I., Paton, N. W. and Mendes, P. *SBRML: a markup language for associating systems biology data with models.* Bioinformatics **26**:932–938 (2010).

[117] Dai, L. *Nonlinear dynamics of piecewise constant systems and implementation of piecewise constant arguments.* World Scientific (2008).

[118] Dangalchev, C. *Residual closeness in networks.* Physica A **365**:556 (2006).

[119] Danon, L., Guilera, D. A., Duch, J. and Arenas, A. *Comparing community structure identification.* Journal of Statistical Mechanics: Theory and Experiment **2005**:P09008 (2005).

[120] Davidich, M. and Bornholdt, S. *The transition from differential equations to Boolean networks: a case study in simplifying a regulatory network model.* J. Theor. Biol. **255**:269–277 (2008).

[121] Davies, J. A. *Developmental biologists' choice of subjects approximates to a power law, with no evidence for the existence of a special group of "model organisms".* BMC Dev. Biol. **7**:40 (2007).

[122] Davis, L. *Handbook of genetic algorithms.* Van Nostrand Reinhold, New York (1991).

[123] de Jong, H., Geiselmann, J., Batt, G., Hernandez, C. and Page, M. *Qualitative simulation of the initiation of sporulation in Bacillus subtilis.* Bull. Math. Biol. **66**:261–299 (2004).

[124] de Jong, H., Geiselmann, J., Hernandez, C. and Page, M. *Genetic Network Analyzer: qualitative simulation of genetic regulatory networks.* Bioinformatics **19**:336–344 (2003).

[125] de Jong, H., Gouze, J. L., Hernandez, C., Page, M., Sari, T. and Geiselmann, J. *Qualitative simulation of genetic regulatory networks using piecewise-linear models.* Bull. Math. Biol. **66**:301–340 (2004).

[126] de Jong, H. and Page, M. *Search for steady states of piecewise-linear differential equation models of genetic regulatory networks.* IEEE/ACM Trans. Comput. Biol. Bioinform. **5**:208–222 (2008).

[127] de Menezes, M. A., Moukarzel, C. and Penna, T. J. P. *First-order transition in small-world networks.* Europhys. Lett. **50**:574–579 (2000).

[128] de Vries, G., Hillen, T., Lewis, M., Müller, J. and Schönfisch, B. *A Course in Mathematical Biology: Quantitative Modeling with Mathematical & Computational Methods.* SIAM (2006).

[129] Deeds, E. J., Ashenberg, O. and Shakhnovich, E. I. *A simple physical model for scaling in protein-protein interaction networks.* PNAS **103**:311–316 (2006).

[130] Degenring, D., Froemel, C., Dikta, G. and Takors, R. *Sensitivity analysis for the reduction of complex metabolism models.* Journal of Process Control **14**:729745 (2004).

[131] Derényi, I., Palla, G. and Vicsek, T. *Clique percolation in random networks.* Phys. Rev. Lett. **94**:160202 (2005).

[132] Deutsch, A. and Dormann, S. *Cellular Automaton Modeling of Biological Pattern Formation.* Birkhäuser Boston (2005).

[133] D'haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. *Linear modeling of mRNA expression levels during CNS development and injury.* Pac. Symp. Biocomput. pages 41–52 (1999).

[134] Dhar, P. K., Meng, T. C., Somani, S., Ye, L., Sakharkar, K., Krishnan, A., Ridwan, A. B., Wah, S. H., Chitre, M. and Hao, Z. *Grid cellware: the first grid-enabled tool for modelling and simulating cellular processes.* Bioinformatics **21**:1284–1287 (2005).

[135] Di Cara, A., Garg, A., De Micheli, G., Xenarios, I. and Mendoza, L. *Dynamic simulation of regulatory networks using* `SQUAD`. BMC Bioinformatics **8**:462 (2007).

[136] Dickinson, R. and Gelinas, R. *Sensitivity analysis of ordinary differential equation systems. A direct method.* J. Comput. Phys. **21**:123–143 (1976).

[137] Do, J. H., Nagasaki, M. and Miyano, S. *The systems approach to the prespore-specific activation of sigma factor SigF in* Bacillus subtilis. Biosystems **100**(3):178–184 (2010).

[138] Doi, A., Fujita, S., Matsuno, H., Nagasaki, M. and Miyano, S. *Constructing biological pathway models with hybrid functional Petri nets.* In Silico Biol. **4**:271–291 (2004).

[139] Doi, A., Nagasaki, M., Matsuno, H. and Miyano, S. *Simulation-based validation of the p53 transcriptional activity with hybrid functional petri net.* In Silico Biol. **6**:1–13 (2006).

[140] Donetti, L. and Muñoz, M. A. *Detecting network communities: a new systematic and efficient algorithm.* Journal of Statistical Mechanics: Theory and Experiment **2004**(10):P10012 (2004).

[141] Dorogovtsev, S. N., Goltsev, A. V. and Mendes, J. F. *Pseudofractal scale-free web.* Phys. Rev. E **65**:066122 (2002).

[142] Dorogovtsev, S. N. and Mendes, J. F. *Evolution of networks with aging of sites.* Phys. Rev. E **62**:1842–1845 (2000).

[143] Doyle, J. C., Alderson, D. L., Li, L., Low, S., Roughan, M., Shalunov, S., Tanaka, R. and Willinger, W. *The "robust yet fragile" nature of the Internet.* PNAS **102**:14497–14502 (2005).

[144] Dräger, A., Planatscher, H., Motsou Wouamba, D., Schröder, A., Hucka, M., Endler, L., Golebiewski, M., Müller, W. and Zell, A. *SBML2LaTeX: conversion of SBML files into human-readable reports.* Bioinformatics **25**:1455–1456 (2009).

[145] Dräger, A. and Hassis, N. and Supper, J. and Schröder, A. and Zell, A. *SBMLsqueezer: a CellDesigner plug-in to generate kinetic rate equations for biochemical networks.* BMC Syst. Biol. **2**:39 (2008).

[146] Duch, J. and Arenas, A. *Community detection in complex networks using extremal optimization.* Phys. Rev. E **72**:027104 (2005).

[147] Dunker, J. *The decoupled direct method for calculating sensitivities coefficients in chemical kinetics.* J. Chem. Phys. **81**:2385–2393 (1984).

[148] Dunn, R., Dudbridge, F. and Sanderson, C. M. *The use of edge-betweenness clustering to investigate biological function in protein interaction networks.* BMC Bioinformatics **6**:39 (2005).

[149] Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J. and Gerstein, M. *Bridging structural biology and genomics: assessing protein interaction data with known complexes.* Trends Genet. **18**:529–536 (2002).

[150] Egea, J., Rodriguez-Fernandez, M., Banga, J. R. and Martí, R. *Scatter Search for chemical and bioprocess optimization.* Journal of Global Optimization **37**(3):481–503 (2007).

[151] Egea, J., Vazquez, E., Banga, J. R. and Martí, R. *Improved scatter search for the global optimization of computationally expensive dynamic models.* Journal of Global Optimization **43**(2-3):175–190 (2009).

[152] Eisenberg, E. and Levanon, E. Y. *Preferential attachment in the protein network evolution.* Phys. Rev. Lett. **91**:138701 (2003).

[153] Eißing, T. and Allgöwer, F. and Bullinger, E. *Robustness properties of apoptosis models with respect to parameter variations and intrinsic noise.* Syst. Biol. **152**:221–228 (2005).

[154] Eißing, T. and Waldherr, S. and Allgöwer, F. and Scheurich, P. and Bullinger, E. *Steady state and (bi-) stability evaluation of simple protease signalling networks.* BioSystems **90**:591–601 (2007).

[155] Elowitz, M. B. and Leibler, S. *A synthetic oscillatory network of transcriptional regulators.* Nature **403**:335–338 (2000).

[156] Elsgolts, L. and Norkin, S. *Introduction to the theory and application of differential equations with deviating arguments.* Academic Press (1973).

[157] Erdős, P. and Rényi, A. *On Random Graphs. I.* Publicationes Mathematicae **6**:290–297 (1959).

[158] Erdős, P. and Rényi, A. *On the evolution of random graphs.* Publ. Math. Inst. Hung. Acad. Sci. **5**:17–61 (1960).

[159] Erdős, P. and Rényi, A. *On the strength of connectedness of a random graph.* Acad. Sci. Hungar. **12**:261–267 (1961).

[160] Ermentrout, G. B. and Edelstein-Keshet, L. *Cellular automata approaches to biological modeling.* J. Theor. Biol. **160**:97–133 (1993).

[161] Fall, P., Marland, E. S., Wagner, J. M. and Tyson, J. J. *Computational Cell Biology.* Springer (2002).

[162] Faller, D., Klingmüller, U. and Timmer, J. *Simulation Methods for Optimal Experimental Design in Systems Biology.* SIMULATION **79**:717–725 (2003).

[163] Farina, M. and Findeisen, R. and Bullinger, E. and Bittanti, S. and Allöwer, F. and Wellstead, P. *Results Towards Identifiability Properties of Biochemical Reaction Networks.* Proc. of the 45th IEEE Conference on Decision & Control pages 2104–2109 (2006).

[164] Farkas, I., Derényi, I., Jeong, H., Néda, Z., Oltvai, Z. N., Ravasz, E., Schubert, A., Barabási and Vicsek, T. *Networks in life: scaling properties and eigenvalue spectra.* Physica A: Statistical Mechanics and its Applications **314**(1-4):25–34 (2002).

[165] Farkas, I. J., Wu, C., Chennubhotia, C., Bahar, I. and Oltvai, Z. N. *Topological basis of signal integration in the transcriptional-regulatory network of the yeast,* Saccharomyces cerevisiae. BMC Bioinformatics **7**:478 (2006).

[166] Farutin, V., Robison, K., Lightcap, E., Dancik, V., Ruttenberg, A., Letovsky, S. and Pradines, J. *Edge-count probabilities for the identification of local protein communities and their organization.* Proteins **62**:800–818 (2006).

[167] Fausett, L. *Neural Networks for Pattern Recognition Fundamentals of Neural Networks: Architectures, Algorithms, and Applications.* Prentice Hall (1994).

[168] Featherstone, D. E. and Broadie, K. *Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network.* Bioessays **24**:267–274 (2002).

[169] Feist, A., Henry, C. S., Reed, J. L., Krummenacker, A., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V. and Palsson, B. Ø. *A genome-scale metabolic reconstruction for* Escherichia coli *K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol. Syst. Biol. **3**:121 (2007).

[170] Fell, D. A. *Metabolic control analysis: a survey of its theoretical and experimental development.* Biochem. J. **286**(2):313–330 (1992).

[171] Fell, D. A. *Understanding the Control of Metabolism.* Portland Press (1997).

[172] Fell, D. A. and Wagner, A. *The small world of metabolism.* Nat. Biotechnol. **18**:1121–1122 (2000).

[173] Feng, X. J., Hooshangi, S., Chen, D., Li, G., Weiss, R. and Rabitz, H. *Optimizing genetic circuits by global sensitivity analysis.* Biophys. J. **87**:2195–2202 (2004).

[174] Feng, X. J. and Rabitz, H. *Optimal identification of biochemical reaction networks.* Biophys. J. **86**:1270–1281 (2004).

[175] Feng, X. J., Rabitz, H., Turinici, G. and Le Bris, C. *A closed-loop identification protocol for nonlinear dynamical systems.* J. Phys. Chem. A **110**:7755–7762 (2006).

[176] Flaherty, P., Jordan, M. I. and Arkin, A. P. *Robust design of biological experiments.* In *Proceeding on the Neural Information Processing Symposium* (2005).

[177] Fomekong-Nanfack, Y., Kaandorp, J. A. and Blom, J. *Efficient parameter estimation for spatio-temporal models of pattern formation: case study of* Drosophla melanogaster. Bioinformatics **23**:3356–3363 (2007).

[178] Fomekong-Nanfack, Y., Postma, M. and Kaandorp, J. A. *Inferring* Drosophila *gap gene regulatory network: a parameter sensitivity and perturbation analysis.* BMC Syst. Biol. **3**:94 (2009).

[179] Fortunato, S. *Community detection in graphs.* Physics Reports **486**:75–174 (2010).

[180] Fortunato, S. and Barthélemy, M. *Resolution limit in community detection.* PNAS **104**:36–41 (2007).

[181] Fortunato, S., Latora, V. and Marchiori, M. *Method to find community structures based on information centrality.* Phys. Rev. E **70**:056104 (2004).

[182] Fraser, H. B. *Modularity and evolutionary constraint on proteins.* Nat. Genet. **37**:351–352 (2005).

[183] Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. and Feldman, M. W. *Evolutionary rate in the protein interaction network.* Science **296**:750–752 (2002).

[184] Freeman, L. C. *A set of measures of centrality based on betweenness.* Sociometry **40**(1):35–41 (1977).

[185] Freeman, M. *Feedback control of intercellular signalling in development.* Nature **408**:313–319 (2000).

[186] Frey, H. C. and Patil, S. R. *Identification and review of sensitivity analysis methods.* Risk Anal. **22**:553–578 (2002).

[187] Friedman, N., Linial, M., Nachman, I. and Pe'er, D. *Using Bayesian networks to analyze expression data.* J. Comput. Biol. **7**:601–620 (2000).

[188] Fullana, J. M. and Rossi, M. *Identification methods for nonlinear stochastic systems.* Phys. Rev. E **65**:031107 (2002).

[189] Funahashi, A., Tanimura, N., Morohashi, M. and Kitano, H. *`CellDesigner`: a process diagram editor for gene-regulatory and biochemical networks.* BIOSILICO **1**:159–162 (2003).

[190] Gadkar, K. G., Gunawan, R. and Doyle, F. J. *Iterative approach to model identification of biological networks.* BMC Bioinformatics **6**:155 (2005).

[191] Gadkar, K. G., Varner, J. and Doyle, F. J. *Model identification of signal transduction networks from data using a state regulator problem.* Syst. Biol. **2**:17–30 (2005).

[192] Gardiner, C. *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences.* Springer (2004).

[193] Gardner, T. S., Cantor, C. R. and Collins, J. J. *Construction of a genetic toggle switch in* Escherichia coli. Nature **403**:339–342 (2000).

[194] Gardónyi, M., Jeppsson, M., Lidén, G., Gorwa-Grauslund, M. F. and Hahn-Hägerdal, B. *Control of xylose consumption by xylose transport in recombinant Saccharomyces cerevisiae.* Biotechnol. Bioeng. **82**:818–824 (2003).

[195] Gautreau, A., Barrat, A. and Barthélemy, M. *Microdynamics in stationary complex networks.* PNAS **106**:8847–8852 (2009).

[196] Gebert, J., Radde, N. and Weber, G.-W. *Modeling gene regulatory networks with piecewise linear differential equations.* European Journal of Operational Research **181**(3):1148–1165 (2007).

[197] Gennemark, P. and Wedelin, D. *Efficient algorithms for ordinary differential equation model identification of biological systems.* IET Syst. Biol. **1**:120–129 (2007).

[198] Gennemark, P. and Wedelin, D. *Benchmarks for identification of ordinary differential equations from time series data.* Bioinformatics **25**:780–786 (2009).

[199] Genrich, H., Kuffner, R. and Voss, K. *Executable petri net models for the analysis of metabolic pathways.* International J. Software Tools for Technology Transfer **3**(4):394–404 (2001).

[200] Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. and Weissman, J. S. *Global analysis of protein expression in yeast.* Nature **425**:737–741 (2003).

[201] Gibson, M. A. and Bruck, J. *Efficient exact stochastic simulation of chemical systems with many species and many channels.* J. Phys. Chem. **104**:1876–1889 (2000).

[202] Gierer, A. and Meinhardt, H. *A theory of biological pattern formation.* Kybernetik **12**:30–39 (1972).

[203] Gilbert, E. *Random Graphs.* Annals of Mathematical Statistics **30**:1141–1144 (1959).

[204] Gillespie, D. T. *Exact stochastic simulation of coupled chemical reactions.* J. Phys. Chem. **81**:2340–2361 (1977).

[205] Gillespie, D. T. *A rigorous derivation of the chemical master equation.* Physica A **188**:404–425 (1992).

[206] Gillespie, D. T. *Approximate accelerated stochastic simulation of chemically reacting systems.* J. Chem. Phys. **115**:1716–1733 (2001).

[207] Girard, S., Luckhoff, A., Lechleiter, J., Sneyd, J. and Clapham, D. *Two-dimensional model of calcium waves reproduces the patterns observed in* Xenopus *oocytes.* Biophys. J. **61**:509–517 (1992).

[208] Girvan, M. and Newman, M. E. *Community structure in social and biological networks.* PNAS **99**:7821–7826 (2002).

[209] Glass, L. *Classification of biological networks by their qualitative dynamics.* J. Theor. Biol. **54**:85–107 (1975).

[210] Glass, L. and Kauffman, S. A. *The logical analysis of continuous, nonlinear biochemical control networks.* J. Theor. Biol. **39**:103–129 (1973).

[211] Glover, F. *Tabu Search—Part I.* ORSA Journal on Computing **1(3)**:190–206 (1989).

[212] Glover, F. *Tabu Search—Part II.* ORSA Journal on Computing **2(1)**:4–32 (1990).

[213] Glover, F. and Laguna, F. *Tabu Search.* Kluwer Academic Publishers (1997).

[214] Goel, G., Chou, I. C. and Voit, E. O. *System estimation from metabolic time-series data.* Bioinformatics **24**:2505–2511 (2008).

[215] Goel, N., Maitra, S. and Montroll, E. *On the Volterra and Other Non-Linear Models of Interacting Populations.* Academic Press (1971).

[216] Goh, K., Oh, E., Jeong, H., Kahng, B. and Kim, D. *Classification of scale-free networks.* PNAS **99**:12583–12588 (2002).

[217] Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley (1989).

[218] Goldbeter, A. *A model for circadian oscillations in the* Drosophila *period protein (PER).* Proc. Biol. Sci. **261**:319–324 (1995).

[219] Goldstein, M., Morris, S. and Yen, G. *Problems with fitting to the power-law distribution.* Eur. Phys. J. B **41**:255–258 (2004).

[220] Gonzalez, O. R. and Küper, C. and Jung, K. and Naval, P. C. and Mendoza, E. . *Parameter estimation using Simulated Annealing for S-system models of biochemical networks.* Bioinformatics **23**:480–486 (2007).

[221] Gonze, D., Halloy, J. and Goldbeter, A. *Robustness of circadian rhythms with respect to molecular noise.* PNAS **99**:673–678 (2002).

[222] Gonze, D., Halloy, J. and Goldbeter, A. *Emergence of coherent oscillations in stochastic models for circadian rhythms.* Physica A **342**:221–233 (2004).

[223] Goss, P. J. and Peccoud, J. *Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets.* PNAS **95**:6750–6755 (1998).

[224] Gourley, S. A., Kuang, Y. and Nagy, J. D. *Dynamics of a delay differential equation model of hepatitis B virus infection.* Journal of Biological Dynamics **2**:140–153 (2008).

[225] Grewal, M. S. and Glover, K. *Identifiability of Linear and Nonlinear Dynamical Systems.* IEEE Transactions on Automatic Control **21**(6):833–837 (1976).

[226] Griepentrog, E. and März, R. *Differential algebraic equations and their numerical solution.* Teubner (1986).

[227] Grigorov, M. G. *Global properties of biological networks.* Drug Discov. Today **10**:365–372 (2005).

[228] Grognard, F., de Jong, H. and Gouzé, J.-L. *Piecewise-Linear Models of Genetic Regulatory Networks: Theory and Example.* Lecture Notes in Control and Information Sciences **357**:137–159 (2007).

[229] Grunwald, S., Speer, A., Ackermann, J. and Koch, I. *Petri net modelling of gene regulation of the Duchenne muscular dystrophy.* BioSystems **92**:189–205 (2008).

[230] Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F. and Arenas, A. *The real communication network behind the formal chart: Community structure in organizations.* Journal of Economic Behavior & Organization **61**(4):653–667 (2006).

[231] Guimerà, R. and Nunes Amaral, L. A. *Functional cartography of complex metabolic networks.* Nature **433**:895–900 (2005).

[232] Guimerà, R., Sales-Pardo, M. and Amaral, L. A. *Modularity from fluctuations in random graphs and complex networks.* Phys. Rev. E **70**:025101 (2004).

[233] Guimerà, R., Sales-Pardo, M. and Amaral, L. A. *A network-based method for target selection in metabolic networks.* Bioinformatics **23**:1616–1622 (2007).

[234] Guimerà, R., Sales-Pardo, M. and Amaral, L. A. *Classes of complex networks defined by role-to-role connectivity profiles.* Nat. Phys. **3**:63–69 (2007).

[235] Gulbahce, N. and Lehmann, S. *The art of community detection.* Bioessays **30**:934–938 (2008).

[236] Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R. and Sethna, J. P. *Universally sloppy parameter sensitivities in systems biology models.* PLoS Comput. Biol. **3**:1871–1878 (2007).

[237] Hahn, M. W. and Kern, A. D. *Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks.* Mol. Biol. Evol. **22**:803–806 (2005).

[238] Hamed, R. I., Ahson, S. I. and Parveen, R. *A new approach for modelling gene regulatory networks using fuzzy petri nets.* J. Integr. Bioinform. **7**(1):113 (2010).

[239] Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P. and Vidal, M. *Evidence for dynamically organized modularity in the yeast protein-protein interaction network.* Nature **430**:88–93 (2004).

[240] Handl, J., Kell, D. B. and Knowles, J. *Multiobjective optimization in bioinformatics and computational biology.* IEEE/ACM Trans. Comput. Biol. Bioinform. **4**:279–292 (2007).

[241] Hasenauer, J., Waldherr, S., Wagner, K. and Allgower, F. *Parameter identification, experimental design and model falsification for biological network models using semidefinite programming.* IET Syst. Biol. **4**:119–130 (2010).

[242] Hastings, M. B. *Community detection as an inference problem.* Phys. Rev. E **74**:035102 (2006).

[243] Hastings, W. K. *Monte Carlo sampling methods using Markov chains and their applications.* Biometrika **57**(1):97–109 (1970).

[244] Hawari, A. H. and Mohamed-Hussein, Z. A. *Simulation of a Petri net-based model of the terpenoid biosynthesis pathway.* BMC Bioinformatics **11**:83 (2010).

[245] Heiner, M., Koch, I. and Will, J. *Model validation of biological pathways using Petri nets–demonstrated for apoptosis.* BioSystems **75**:15–28 (2004).

[246] Heinrich, R. and Rapoport, T. A. *A linear steady-state treatment of enzymatic chains. Critique of the crossover theorem and a general procedure to identify interaction sites with an effector.* Eur. J. Biochem. **42**:97–105 (1974).

[247] Heinrich, R. and Rapoport, T. A. *A linear steady-state treatment of enzymatic chains. General properties, control and effector strength.* Eur. J. Biochem. **42**:89–95 (1974).

[248] Heinrich, R. and Schuster, S. *The Regulation of Cellular Systems.* Chapman and Hall (1996).

[249] Hengl, S., Kreutz, C., Timmer, J. and Maiwald, T. *Data-based identifiability analysis of non-linear dynamical models.* Bioinformatics **23** (2007).

[250] Higgins, J. *Analysis of sequential reactions.* Ann. N. Y. Acad. Sci. **108**:305–321 (1963).

[251] Higham, C. F. *Bifurcation analysis informs Bayesian inference in the Hes1 feedback loop.* BMC Syst. Biol. **3**:12 (2009).

[252] Higham, D. J., Rasajski, M. and Pržulj, N. *Fitting a geometric graph to a protein-protein interaction network.* Bioinformatics **24**:1093–1099 (2008).

[253] Hindmarsh, A. C. *LSODE and LSODI, two new initial value ordinary differential equation solvers.* ACM SIGNUM Newsletter **15**(4):10–11 (1980).

[254] Hindmarsh, A. C., Brown, P. N., Grant, K. E., Lee, S. L., Serban, R., Shumaker, D. E. and Woodward, C. S. *SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers.* ACM Transactions on Mathematical Software **31**(3):363–396 (2005).

[255] Hirmajer, T., Balsa-Canto, E. and Banga, J. R. *DOTcvpSB, a software toolbox for dynamic optimization in systems biology.* BMC Bioinformatics **10**:199 (2009).

[256] Hodgkin, A. L. and Huxley, A. F. *A quantitative description of membrane current and its application to conduction and excitation in nerve.* J. Physiol. **117**:500–544 (1952).

[257] Hoefnagel, M. H., Starrenburg, M. J., Martens, D. E., Hugenholtz, J., Kleerebezem, M., Van Swam, I. I., Bongers, R., Westerhoff, H. V. and Snoep, J. L. *Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis.* Microbiology **148**:1003–1013 (2002).

[258] Höfer, T., Nathansen, H., Lohning, M., Radbruch, A. and Heinrich, R. *GATA-3 transcriptional imprinting in Th2 lymphocytes: a mathematical model.* PNAS **99**:9364–9368 (2002).

[259] Hofmeyr, J. H. S. *Metabolic control analysis in a nutshell.* Proceedings of the 2nd International Conference on Systems Biology pages 291–300 (2001).

[260] Holland, J. *Adaptation in Natural and Artificial Systems.* University of Michigan Press (1975).

[261] Holme, P., Huss, M. and Jeong, H. *Subnetwork hierarchies of biochemical pathways.* Bioinformatics **19**:532–538 (2003).

[262] Hoogenboom, J., den Otter, W. and Offerhaus, H. *Accurate and unbiased estimation of power-law exponents from single-emitter blinking data.* The Journal of Chemical Physics **125**:204713 (2006).

[263] Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. and Kummer, U. *COPASI–a COmplex PAthway SImulator.* Bioinformatics **22**:3067–3074 (2006).

[264] Hu, D. and Yuan, J. M. *Time-dependent sensitivity analysis of biological networks: coupled MAPK and PI3K signal transduction pathways.* J. Phys. Chem. A **110**:5361–5370 (2006).

[265] Hu, Y., Chen, H., Zhang, P., Li, M., Di, Z. and Fan, Y. *Comparative definition of community and corresponding identifying algorithm.* Phys. Rev. E **78**:026121 (2008).

[266] Hu, Y., Li, M., Zhang, P., Fan, Y. and Di, Z. *Community detection by signaling on complex networks.* Phys. Rev. E **78**:016115 (2008).

[267] Huang, S. *Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery.* J. Mol. Med. **77**:469–480 (1999).

[268] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H. and the rest of the SBML Forum. *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.* Bioinformatics **19**:524–531 (2003).

[269] Humphries, M. D. and Gurney, K. *Network "small-world-ness": a quantitative method for determining canonical network equivalence.* PLoS ONE **3**:e0002051 (2008).

[270] Ingalls, B. *A Frequency Domain Approach to Sensitivity Analysis of Biochemical Networks.* J. Phys. Chem. B **108**:1143–1152 (2004).

[271] Ingalls, B. *Sensitivity analysis: from model parameters to system behaviour.* Essays Biochem. **45**:177–193 (2008).

[272] Ironi, L. and Panzeri, L. *A computational framework for qualitative simulation of nonlinear dynamical models of gene-regulatory networks.* BMC Bioinformatics **10**(Suppl 12):S14 (2009).

[273] Izaguirre, J. A., Chaturvedi, R., Huang, C., Cickovski, T., Coffland, J., Thomas, G., Forgacs, G., Alber, M., Hentschel, G., Newman, S. A. and Glazier, J. A. *`CompuCell`, a multi-model framework for simulation of morphogenesis.* Bioinformatics **20**:1129–1137 (2004).

[274] Jacquez, J. A. and Perry, T. *Parameter estimation: local identifiability of parameters.* Am. J. Physiol. **258**:E727–E736 (1990).

[275] Jaeger, J., Blagov, M., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Surkova, S., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H. and Reinitz, J. *Dynamical Analysis of Regulatory Interactions in the Gap Gene System of* Drosophila melanogaster. Genetics **167**:1721–1737 (2004).

[276] Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Vanario-Alonso, C., Samsonova, M., Sharp, D. H. and Reinitz, J. *Dynamic control of positional information in the early* Drosophila *embryo.* Nature **430**:368–371 (2004).

[277] Janson, S., Łuczak, T. and Rucinski, A. *Random Graphs.* John Wiley, New York (2000).

[278] Jaqaman, K. and Danuser, G. *Linking data to models: data regression.* Nat. Rev. Mol. Cell Biol. **7**:813–819 (2006).

[279] Jeong, H., Mason, S., Barabási, A. and Oltvai, Z. *Lethality and centrality in protein networks.* Nature **411**:41–42 (2001).

[280] Jeong, H., Tombor, R., Albert, R., Oltvai, Z. and Barabási, A.-L. *The large-scale organization of metabolic networks.* Nature **407**:651–654 (2000).

[281] Ji, X. and Xu, Y. *libSRES: a C library for stochastic ranking evolution strategy for parameter estimation.* Bioinformatics **22**:124–126 (2006).

[282] Jones, D. S. and Sleeman, B. D. *Differential Equations and Mathematical Biology.* Chapman & Hall/CRC (2003).

[283] Joshi, M., Seidel-Morgenstern, A. and Kremling, A. *Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems.* Metab. Eng. **8**:447–455 (2006).

[284] Jostins, L. and Jaeger, J. *Reverse engineering a gene network using an asynchronous parallel evolution strategy.* BMC Syst. Biol. **4**:17 (2010).

[285] Joy, M. P., Brock, A., Ingber, D. E. and Huang, S. *High-betweenness proteins in the yeast protein interaction network.* J. Biomed. Biotechnol. **2005**:96–103 (2005).

[286] Junker, B. H. and Schreiber, F. *Analysis of Biological Networks.* Wiley-Interscience (2008).

[287] Kacser, H. and Burns, J. A. *The control of flux.* Symp. Soc. Exp. Biol. **27**:65–104 (1973).

[288] Kacser, H. and Burns, J. A. *The control of flux.* Biochem. Soc. Trans. **23**:341–366 (1995).

[289] Kalir, S., Mangan, S. and Alon, U. *A coherent feed-forward loop with a SUM input function prolongs flagella expression in* Escherichia coli. Mol. Syst. Biol. **1**:2005.0006 (2005).

[290] Karonski, M. *A review of random graphs.* Journal of Graph Theory **6**:349–389 (1982).

[291] Karrer, B., Levina, E. and Newman, M. E. *Robustness of community structure in networks.* Phys. Rev. E **77**:046119 (2008).

[292] Kauffman, S. A. *Metabolic stability and epigenesis in randomly constructed genetic nets.* J. Theor. Biol. **22**:437–467 (1969).

[293] Kauffman, S. A. *The Origins of Order, Self-Organization and Selection in Evolution.* Oxford University Press (1993).

[294] Kernighan, B. W. and Lin, S. *An efficient heuristic procedure for partitioning graphs.* Bell System Technical Journal **49**(1):291–307 (1970).

[295] Keseler, I. M., Bonavides-Martínez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A. G. and Karp, P. D. *EcoCyc: a comprehensive view of* Escherichia coli *biology.* Nucleic Acids Res. **37**:D464–D470 (2009).

[296] Khanin, R. and Wit, E. *How scale-free are biological networks?* J. Comput. Biol. **13**:810–818 (2006).

[297] Kierzek, A. M. *`STOCKS`: STOChastic Kinetic Simulations of biochemical systems with Gillespie algorithm.* Bioinformatics **18**:470–481 (2002).

[298] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K. and Tomita, M. *Dynamic modeling of genetic networks using genetic algorithm and S-system.* Bioinformatics **19**:643–650 (2003).

[299] Kim, J., Krapivsky, P. L., Kahng, B. and Redner, S. *Infinite-order percolation and giant fluctuations in a protein interaction network.* Phys. Rev. E **66**:055101 (2002).

[300] Kim, K. A., Spencer, S. L., Albeck, J. G., Burke, J. M., Sorger, P. K., Gaudet, S. and Kim, D. H. *Systematic calibration of a cell signaling network model.* BMC Bioinformatics **11**:202 (2010).

[301] Kim, P. S., Levy, D. and Lee, P. P. *Modeling and simulation of the immune system as a self-regulating network.* Meth. Enzymol. **467**:79–109 (2009).

[302] Kim, W. K. and Marcotte, E. M. *Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence.* PLoS Comput. Biol. **4**:e1000232 (2008).

[303] Kimura, S., Hatakeyama, M. and Konagaya, A. *Inference of S-system Models of Genetic Networks from Noisy Time-series Data.* Chem-Bio Informatics Journal **4**(1):1–14 (2004).

[304] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. *Optimization by simulated annealing.* Science **220**:671–680 (1983).

[305] Kitano, H. *Towards a theory of biological robustness.* Mol. Syst. Biol. **3**:137 (2007).

[306] Klamt, S., Saez-Rodriguez, J. and Gilles, E. D. *Structural and functional analysis of cellular networks with `CellNetAnalyzer`.* BMC Syst. Biol. **1**:2 (2007).

[307] Klemm, K. and Eguíluz, V. M. *Highly clustered scale-free networks.* Phys. Rev. E **65**:036123 (2002).

[308] Klipp, E., Herwig, R., Kowald, A., Wierling, C. and Lehrach, H. *Systems Biology in Practice.* Wiley-VCH (2005).

[309] Koda, M., McRae, G. and Seinfeld, J. *Automatic sensitivity analysis of kinetic mechanisms.* International Journal of Chemical Kinetics **11**:427–444 (1979).

[310] Kontoravdi, C., Asprey, S. P., Pistikopoulos, E. N. and Mantalaris, A. *Application of global sensitivity analysis to determine goals for design of experiments: an example study on antibody-producing cell cultures.* Biotechnol. Prog. **21**:1128–1135 (2005).

[311] Krauss, S. and Brand, M. D. *Quantitation of signal transduction.* FASEB J. **14**:2581–2588 (2000).

[312] Kravtsov, V., Carmeli, D., Dubitzky, W., Orda, A., Schuster, A., Silberstein, M. and Yoshpa, B. *Quasi-opportunistic Supercomputing in Grid Environments.* Proc. of Intl. Conference on Algorithms and Architectures (ICA3PP 2008) pages 233–244 (2008).

[313] Kremling, A., Fischer, S., Gadkar, K., Doyle, F. J., Sauter, T., Bullinger, E., Allgower, F. and Gilles, E. D. *A benchmark for methods in reverse engineering and model discrimination: problem formulation and solutions.* Genome Res. **14**:1773–1785 (2004).

[314] Kreutz, C. and Timmer, J. *Systems biology: experimental design.* FEBS J. **276**:923–942 (2009).

[315] Kriete, A. and Eils, R. *Computational Systems Biology.* Elsevier Academic Press (2006).

[316] Kuchaiev, O. and Pržulj, N. *Learning the structure of protein-protein interaction networks.* Pac. Symp. Biocomput. **14**:39–50 (2009).

[317] Kuchaiev, O., Rašajski, M., Higham, D. J. and Pržulj, N. *Geometric De-noising of Protein-Protein Interaction Networks.* PLoS Comput. Biol. **5(8)**:e1000454 (2009).

[318] Kutalik, Z., Cho, K. H. and Wolkenhauer, O. *Optimal sampling time selection for parameter estimation in dynamic pathway modeling.* BioSystems **75**:43–55 (2004).

[319] Lancichinetti, A., Fortunato, S. and Radicchi, F. *Benchmark graphs for testing community detection algorithms.* Phys. Rev. E **78**:046110 (2008).

[320] Langtangen, H. P. and Tveito, A. *Advanced Topics in Computational Partial Differential Equations, Numerical Methods and Diffpack Programming.* Springer-Verlag (2003).

[321] Latora, V. and Marchiori, M. *Efficient behavior of small-world networks.* Phys. Rev. Lett. **87**:198701 (2001).

[322] Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J. L. and Hucka, M. *BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular models.* Nucleic Acids Research **34**:D689–D691 (2006).

[323] Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., Crampin, E. J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J. L., Spence, H. D. and Wanner, B. L. *Minimum information requested in the annotation of biochemical models (MIRIAM).* Nat. Biotechnol. **23**:1509–1515 (2005).

[324] Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villeger, A., Boyd, S. E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T. C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D. B., Sander, C., Sauro, H., Snoep, J. L., Kohn, K. and Kitano, H. *The Systems Biology Graphical Notation.* Nat. Biotechnol. **27**:735–741 (2009).

[325] Lee, D. Y., Zimmer, R., Lee, S. Y. and Park, S. *Colored Petri net modeling and simulation of signal transduction pathways.* Metab. Eng. **8**:112–122 (2006).

[326] Lee, Y., Amaral, L. A. N., Canning, D., Meyer, M. and Stanley, H. E. *Universal features in the growth dynamics of complex organizations.* Phys. Rev. Lett. **81**:3275–3278 (1998).

[327] Lehmann, S. and Hansen, L. K. *Deterministic modularity optimization.* European Physical Journal B **60**:83–88 (2007).

[328] Leis, J. and Kramer, M. A. *The Simultaneous Solution and Sensitivity Analysis of Systems Described by Ordinary Differential Equations* . ACM Transactions on Mathematical Software, **14**:45–60 (1988).

[329] Lenbury, Y. and Pornsawad, P. *A delay-differential equation model of the feedback-controlled hypothalamus-pituitary-adrenal axis in humans.* Math. Med. Biol. **22**:15–33 (2005).

[330] Lewis, A. C. F., Jones, N. S., Porter, M. A. and Deane, C. M. *The Function of Communities in Protein Interaction Networks.* http://arxiv.org/abs/0904.0989.

[331] Li, C., Ge, Q. W., Nakata, M., Matsuno, H. and Miyano, S. *Modelling and simulation of signal transductions in an apoptosis pathway by using timed Petri nets.* J. Biosci. **32**:113–127 (2007).

[332] Li, C., Nagasaki, M., Ueno, K. and Miyano, S. *Simulation-based model checking approach to cell fate specification during Caenorhabditis elegans vulval development by hybrid functional Petri net with extension.* BMC Syst. Biol. **3**:42 (2009).

[333] Li, C., Suzuki, S., Ge, Q. W., Nakata, M., Matsuno, H. and Miyano, S. *Structural modeling and analysis of signaling pathways based on Petri nets.* J. Bioinform. Comput. Biol. **4**:1119–1140 (2006).

[334] Li, J., Kuang, Y. and Mason, C. C. *Modeling the glucose-insulin regulatory system and ultradian insulin secretory oscillations with two explicit time delays.* J. Theor. Biol. **242**:722–735 (2006).

[335] Li, L., Alderson, D., Tanaka, R., Doyle, J. C. and Willinger, W. *Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications.* Internet Mathematics **2(4)**:431–523 (2005).

[336] Li, M., Wang, J. and Chen, J. *A Graph-Theoretic Method for Mining Overlapping Functional Modules in Protein Interaction Networks.* Lecture Notes in Computer Science **4983**:208–219 (2008).

[337] Liang, S., Fuhrman, S. and Somogyi, R. *REVEAL, a general reverse engineering algorithm for inference of genetic network architectures.* Pacific Symposium on Biocomputing **3**:18–29 (1998).

[338] Liebermeister, W. *Response to temporal parameter fluctuations in biochemical networks.* J. Theor. Biol. **234**:423–438 (2005).

[339] Lillacci, G. and Khammash, M. *Parameter estimation and model selection in computational biology.* PLoS Comput. Biol. **6**:e1000696 (2010).

[340] Lima-Mendez, G. and van Helden, J. *The powerful law of the power law and other myths in network biology.* Mol Biosyst **5**:1482–1493 (2009).

[341] Lin, W. H., Liu, W. C. and Hwang, M. J. *Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein-protein interaction networks.* BMC Syst. Biol. **3**:32 (2009).

[342] Liu, G., Swihart, M. T. and Neelamegham, S. *Sensitivity, principal component and flux analysis applied to signal transduction: the case of epidermal growth factor mediated signaling.* Bioinformatics **21**:1194–1202 (2005).

[343] Liu, W. C., Lin, W. H., Davis, A. J., Jordán, F., Yang, H. T. and Hwang, M. J. *A network perspective on the topological importance of enzymes and their phylogenetic conservation.* BMC Bioinformatics **8**:121 (2007).

[344] Ljung, L. *System Identification: Theory for the User.* Prentice-Hall (1999).

[345] Lloyd, C. M., Halstead, M. D. and Nielsen, P. F. *CellML: its future, present and past.* Prog. Biophys. Mol. Biol. **85**:433–450 (2004).

[346] Lloyd, C. M., Lawson, J. R., Hunter, P. J. and Nielsen, P. F. *The CellML Model Repository.* Bioinformatics **24**:2122–2123 (2008).

[347] Lodish, H. and Darnell, J. E. *Molecular Cell Biology.* W.H. Freeman & Co. Ltd. 5th rev. edition (2003).

[348] Lotka, A. J. *Analytical Note on Certain Rhythmic Relations in Organic Systems.* PNAS **6**:410–415 (1920).

[349] Lotka, A. J. *Elements of Physical Biology.* Williams and Wilkins (1925).

[350] Lotka, A. *Contribution to the Theory of Periodic Reaction.* J. Phys. Chem. **14 (3)**:271274 (1910).

[351] Lüdtke, N., Panzeri, S., Brown, M., Broomhead, D. S., Knowles, J., Montemurro, M. A. and Kell, D. B. *Information-theoretic sensitivity analysis: a general method for credit assignment in complex networks.* J. R. Soc. Interface **5**:223–235 (2008).

[352] Luo, F., Yang, Y., Chen, C. F., Chang, R., Zhou, J. and Scheuermann, R. H. *Modular organization of protein interaction networks.* Bioinformatics **23**:207–214 (2007).

[353] Ma, H. and Zeng, A. P. *Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.* Bioinformatics **19**:270–277 (2003).

[354] Machina, A., Ponosov, A. and Voit, E. O. *Automated piecewise power-law modeling of biological systems.* J. Biotechnol. (2010).

[355] MacIsaac, K., Wang, T., Gordon, B., Gifford, D., Stormo, G. and Fraenkel, E. *An improved map of conserved regulatory sites for* Saccharomyces cerevisiae. BMC Bioinformatics **7**:113 (2006).

[356] Mahdavi, A., Davey, R. E., Bhola, P., Yin, T. and Zandstra, P. W. *Sensitivity analysis of intracellular signaling pathway kinetics predicts targets for stem cell fate control.* PLoS Comput. Biol. **3**:e130 (2007).

[357] Maini, P. K. *Applications of mathematical modelling to biological pattern formation.* Lecture Notes in Physics. Coherent Structures in Complex Systems **567**:205–217 (2001).

[358] Mandel, J. J., Fuss, H., Palfreyman, N. M. and Dubitzky, W. *Modeling biochemical transformation processes and information processing with* `Narrator`. BMC Bioinformatics **8**:103 (2007).

[359] Mangan, S. and Alon, U. *Structure and function of the feed-forward loop network motif.* PNAS **100**:11980–11985 (2003).

[360] Mangan, S., Itzkovitz, S., Zaslaver, A. and Alon, U. *The incoherent feed-forward loop accelerates the response-time of the* gal *system of* Escherichia coli. J. Mol. Biol. **356**:1073–1081 (2006).

[361] Mangan, S., Zaslaver, A. and Alon, U. *The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks.* J. Mol. Biol. **334**:197–204 (2003).

[362] Manu, Surkova, S., Spirov, A., Gursky, V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C., Sharp, D., Samsonova, M. and Reinitz, J. *Canalization of Gene Expression and Domain Shifts in the* Drosophila *Blastoderm by Dynamical Attractors.* PLoS Comput. Biol. **5**(3):e1000303 (2009).

[363] Maraziotis, I. A., Dimitrakopoulou, K. and Bezerianos, A. *An in silico method for detecting overlapping functional modules from composite biological networks.* BMC Syst. Biol. **2**:93 (2008).

[364] Marino, S., Hogue, I. B., Ray, C. J. and Kirschner, D. E. *A methodology for performing global uncertainty and sensitivity analysis in systems biology.* J. Theor. Biol. **254**:178–196 (2008).

[365] Maslov, S. and Sneppen, K. *Specificity and stability in topology of protein networks.* Science **296**:910–913 (2002).

[366] Matsuno, H., Doi, A., Nagasaki, M. and Miyano, S. *Hybrid Petri net representation of gene regulatory network.* Pac. Symp. Biocomput. pages 341–352 (2000).

[367] Matsuno, H., Tanaka, Y., Aoshima, H., Doi, A., Matsui, M. and Miyano, S. *Biopathways representation and simulation on hybrid functional Petri net.* In Silico Biol. **3**:389–404 (2003).

[368] May, R. M. *Simple mathematical models with very complicated dynamics.* Nature **261**:459–467 (1976).

[369] McAdams, H. H. and Arkin, A. *Stochastic mechanisms in gene expression.* PNAS **94**:814–819 (1997).

[370] McGuffee, S. R. and Elcock, A. H. *Diffussion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm.* PLoS Comput. Biol. **6**:e1000694 (2010).

[371] McKay, M., Beckman, R. and Conover, W. *A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code.* Technometrics **21**(2):239–245 (1979).

[372] McKean, H. P. *Nagumo's equation.* Advances in Math. **6**:209–223 (1970).

[373] McRae, G. J., Tilden, J. W. and Seinfeld, J. H. *Global sensitivity analysis - a computational implementation of the fourier amplitude sensitivity test (FAST).* Computers and Chemical Engineering **6**(1):15–25 (1982).

[374] Medus, A. D. and Dorso, C. O. *Alternative approach to community detection in networks.* Phys. Rev. E **79**:066111 (2009).

[375] Mei, J., He, S., Shi, G., Wang, Z. and Li, W. *Revealing network communities through modularity maximization by a contraction-dilation method.* New Journal of Physics **11**(4):043025 (2009).

[376] Meinhardt, H. *A model of pattern formation in insect embryogenesis.* J. Cell. Sci. **23**:177–139 (1977).

[377] Meinhardt, H. and Gierer, A. *Pattern formation by local self-activation and lateral inhibition.* Bioessays **22**:753–760 (2000).

[378] Mendes, P., Hoops, S., Sahle, S., Gauges, R., Dada, J. and Kummer, U. *Computational modeling of biochemical networks using* `COPASI`. Methods Mol. Biol. **500**:17–59 (2009).

[379] Mendes, P. and Kell, D. *Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation.* Bioinformatics **14**:869–883 (1998).

[380] Merton, R. K. *The Matthew effect in science.* Science **159**:56–63 (1968).

[381] Meshkat, N., Eisenberg, M. and Distefano, J. J. *An algorithm for finding globally identifiable parameter combinations of nonlinear ODE models using Gröbner Bases.* Math. Biosci. **222**:61–72 (2009).

[382] Mete, M., Tang, F., Xu, X. and Yuruk, N. *A structural approach for finding functional modules from large biological networks.* BMC Bioinformatics **9 Suppl 9**:S19 (2008).

[383] Middendorf, M., Ziv, E. and Wiggins, C. H. *Inferring network mechanisms: the* Drosophila melanogaster *protein interaction network.* PNAS **102**:3192–3197 (2005).

[384] Milenković, T., Lai, J. and Pržulj, N. *GraphCrunch: a tool for large network analyses.* BMC Bioinformatics **9**:70 (2008).

[385] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. and Alon, U. *Superfamilies of evolved and designed networks.* Science **303**:1538–1542 (2004).

[386] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. *Network motifs: simple building blocks of complex networks.* Science **298**:824–827 (2002).

[387] Minnhagen, P. and Bernhardsson, S. *The blind watchmaker network: scale-freeness and evolution.* PLoS ONE **3**:e1690 (2008).

[388] Mitchell, M. *An Introduction to Genetic Algorithms.* MIT Press (1996).

[389] Mjolsness, E., Sharp, D. H. and Reinitz, J. *A connectionist model of development.* J. Theor. Biol. **152**:429–453 (1991).

[390] Moles, C. G., Mendes, P. and Banga, J. R. *Parameter estimation in biochemical pathways: a comparison of global optimization methods.* Genome Res. **13**:2467–2474 (2003).

[391] Montañez, R., Medina, M. A., Solé, R. V. and Rodriguez-Caso, C. *When metabolism meets topology: Reconciling metabolite and reaction networks.* Bioessays **32**:246–256 (2010).

[392] Moraru, I. I., Schaff, J. C., Slepchenko, B. M., Blinov, M. L., Morgan, F., Lakshminarayana, A., Gao, F., Li, Y. and Loew, L. M. *Virtual Cell modelling and simulation software environment.* IET Syst. Biol. **2**:352–362 (2008).

[393] Morohashi, M., Winn, A. E., Borisuk, M. T., Bolouri, H., Doyle, J. and Kitano, H. *Robustness as a measure of plausibility in models of biochemical networks.* J. Theor. Biol. **216**:19–30 (2002).

[394] Morris, M. *Factorial sampling plans for preliminary computational experiments.* Technometrics, **33**(3):161–174 (1991).

[395] Moukarzel, C. F. *Spreading and shortest paths in systems with sparse long-range connections.* Phys. Rev. E **60**:R6263–6266 (1999).

[396] Muff, S., Rao, F. and Caflisch, A. *Local modularity measure for network clusterizations.* Phys. Rev. E **72**:056107 (2005).

[397] Mura, I. and Csikasz-Nagy, A. *Stochastic Petri Net extension of a yeast cell cycle model.* J. Theor. Biol. **254**:850–860 (2008).

[398] Muraille, E., Thieffry, D., Leo, O. and Kaufman, M. *Toxicity and neuroendocrine regulation of the immune response: a model analysis.* J. Theor. Biol. **183**:285–305 (1996).

[399] Murray, J. *Mathematical Biology I and II.* Phys. Rev. E **72**:056107 (2005).

[400] Nagasaki, M., Doi, A., Matsuno, H. and Miyano, S. *A versatile petri net based architecture for modeling and simulation of complex biological processes.* Genome Inform. **15**:180–197 (2004).

[401] Nagasaki, M., Doi, A., Matsuno, H. and Miyano, S. *Computational Modeling of Biological Processes with Petri Net-Based Architecture.* Bioinformatics Technologies (Springer) pages 179–242 (2005).

[402] Nelson, P. W. and Perelson, A. S. *Mathematical analysis of delay differential equation models of HIV-1 infection.* Math Biosci **179**:73–94 (2002).

[403] Newman, M. E. *Renormalization group analysis of the small-world network model.* Physics Letters A **263**(4-6):341–346 (1999).

[404] Newman, M. E. *Models of the Small World.* Journal of Statistical Physics **101**:819–841 (2000).

[405] Newman, M. E. *Assortative mixing in networks.* Phys. Rev. Lett. **89**:208701 (2002).

[406] Newman, M. E. *Mixing patterns in networks.* Phys. Rev. E **67**:026126 (2003).

[407] Newman, M. E. *The structure and function of complex networks.* SIAM Review **45**:167 (2003).

[408] Newman, M. E. *Detecting community structure in networks.* Eur. Phys. J. B **38**:321–330 (2004).

[409] Newman, M. E. *Fast algorithm for detecting community structure in networks.* Phys. Rev. E **69**:066133 (2004).

[410] Newman, M. E. *A measure of betweenness centrality based on random walks.* Social Networks **27**:39–54 (2005).

[411] Newman, M. E. *Power laws, Pareto distributions and Zipf's law.* Contemporary Physics **46(5)**:323–351 (2005).

[412] Newman, M. E. *Finding community structure in networks using the eigenvectors of matrices.* Phys. Rev. E **74**:036104 (2006).

[413] Newman, M. E. *Modularity and community structure in networks.* PNAS **103**:8577–8582 (2006).

[414] Newman, M. E. and Girvan, M. *Finding and evaluating community structure in networks.* Phys. Rev. E **69**:026113 (2004).

[415] Newman, M. E. and Leicht, E. A. *Mixture models and exploratory analysis in networks.* PNAS **104**:9564–9569 (2007).

[416] Newman, M. E., Strogatz, S. H. and Watts, D. J. *Random graphs with arbitrary degree distributions and their applications.* Phys. Rev. E **64**:026118 (2001).

[417] Newman, M. E. and Watts, D. J. *Scaling and percolation in the small-world network model.* Phys. Rev. E **60**:7332–7342 (1999).

[418] Nielsen, J. N., Madsen, H. and Young, P. *Parameter estimation in stochastic differential equations: An overview.* Annu. Rev. Control **24**:8394 (2000).

[419] Noh, J. D. and Rieger, H. *Random walks on complex networks.* Phys. Rev. Lett. **92**:118701 (2004).

[420] Novák, B., Pataki, Z., Ciliberto, A. and Tyson, J. J. *Mathematical model of the cell division cycle of fission yeast.* Chaos **11**:277–286 (2001).

[421] Øksendal, B. *Stochastic differential equations: an introduction with applications.* Springer (2003).

[422] Olivier, B. G. and Snoep, J. L. *Web-based kinetic modelling using JWS Online.* Bioinformatics **20**:2143–2144 (2004).

[423] Ozana, M. *Incipient spanning cluster on small-world networks.* Europhys. Lett. **55**:762–766 (2001).

[424] Palla, G., Derényi, I., Farkas, I. and Vicsek, T. *Uncovering the overlapping community structure of complex networks in nature and society.* Nature **435**:814–818 (2005).

[425] Palla, G., Farkas, I., Derényi, I., Pollner, P. and Vicsek, T. *Directed network modules.* New Journal of Physics **9**:186 (2007).

[426] Park, L. J., Park, C. H., Park, C. and Lee, T. *Application of genetic algorithms to parameter estimation of bioprocesses.* Med. Biol. Eng. Comput. **35**:47–49 (1997).

[427] Pastor-Satorras, R., Smith, E. and Solé, R. *Evolving protein interaction networks through gene duplication.* J. Theor. Biol. **222**:199–210 (2003).

[428] Peifer, M. and Timmer, J. *Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting.* IET Syst. Biol. **1**:78–88 (2007).

[429] Peleg, M., Rubin, D. and Altman, R. B. *Using Petri Net tools to study properties and dynamics of biological systems.* J. Am. Med. Inform. Assoc. **12**:181–199 (2005).

[430] Peleg, M., Yeh, I. and Altman, R. B. *Modelling biological processes using workflow and Petri Net models.* Bioinformatics **18**:825–837 (2002).

[431] Penrose, M. D. *Random Geometric Graphs.* Oxford University Press (2003).

[432] Pereira-Leal, J. B., Enright, A. J. and Ouzounis, C. A. *Detection of functional modules from protein interaction networks.* Proteins **54**:49–57 (2004).

[433] Perkins, T. J., Jaeger, J., Reinitz, J. and Glass, L. *Reverse engineering the gap gene network of* Drosophila melanogaster. PLoS Comput. Biol. **2**:e51 (2006).

[434] Peterson, J. *Petri Nets.* Computing Surveys **9**(3):223–252 (1977).

[435] Peterson, J. *Petri Net Theory and the Modeling of Systems.* Prentice Hall (1981).

[436] Petzold, L. *Differential-Algebraic Equations are not ODE's.* SIAM J. on Scientific and Statistical Computing **3**:367–384 (1982).

[437] Pierce, T. and Cukier, R. *Global nonlinear sensitivity analysis using walsh functions.* Journal of Computational Physics **41**(2):427–443 (1981).

[438] Pinney, J. W., Westhead, D. R. and McConkey, G. A. *Petri Net representations in systems biology.* Biochem. Soc. Trans. **31**:1513–1515 (2003).

[439] Pisarev, A., Poustelnikova, E., Samsonova, M. and Reinitz, J. *FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution.* Nucleic Acids Res. **37**:D560–D566 (2009).

[440] Polisetty, P. K., Voit, E. O. and Gatzke, E. P. *Identification of metabolic system parameters using global optimization methods.* Theor. Biol. Med. Model. **3**:4 (2006).

[441] Polynikis, A., Hogan, S. J. and di Bernardo, M. *Comparing different ODE modelling approaches for gene regulatory networks.* J. Theor. Biol. **261**:511–530 (2009).

[442] Porter, M. A., Onnela, J.-P. and Mucha, P. J. *Communities in Networks.* Notices of the American Mathematical Society **56**:1082–1097 (2009).

[443] Pržulj, N. *Geometric Local Structure in Biological Networks.* Proceedings of the 2007 IEEE Information Theory Workshop (2007).

[444] Price, D. J. *Networks of scientific papers.* Science **149**:510–515 (1965).

[445] Price, D. J. *A general theory of bibliometric and other cumulative advantage processes.* Journal of the American Society for Information Science **27**(5):292–306 (1976).

[446] Pržulj, N. *Biological network comparison using graphlet degree distribution.* Bioinformatics **23**:e177–e183 (2006).

[447] Pržulj, N., Corneil, D. G. and Jurisica, I. *Modeling interactome: scale-free or geometric?* Bioinformatics **20**:35083515 (2004).

[448] Pržulj, N. and Higham, D. J. *Modelling protein-protein interaction networks via a stickiness index.* J. R. Soc. Interface **3(10)**:711–716 (2006).

[449] Pržulj, N., Kuchaiev, O., Stevanović, A. and Hayes, W. *Geometric evolutionary dynamics of protein interaction networks.* Pac. Symp. Biocomput. pages 178–189 (2010).

[450] Pujol, J. M., Béjar, J. and Delgado, J. *Clustering algorithm for determining community structure in large networks.* Phys. Rev. E **74**:016107 (2006).

[451] Qian, J., Luscombe, N. M. and Gerstein, M. *Protein Family and Fold Occurrence in Genomes: Power-law Behavior and Evolutionary Model.* J. Mol. Biol **313**:673–681 (2001).

[452] Quaiser, T. and Monnigmann, M. *Systematic identifiability testing for unambiguous mechanistic modeling–application to JAK-STAT, MAP kinase, and NF-kappaB signaling pathway models.* BMC Syst. Biol. **3**:50 (2009).

[453] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D. *Defining and identifying communities in networks.* PNAS **101**:2658–2663 (2004).

[454] Raghavan, U. N., Albert, R. and Kumara, S. *Near linear time algorithm to detect community structures in large-scale networks.* Phys. Rev. E **76**:036106 (2007).

[455] Ramsay, J. O., Hooker, G., Campbell, D., Cao, J. and Ramsay, J. O. *Parameter estimation for differential equations: A generalized smoothing approach.* Journal of the Royal Statistical Society **69**(5):741–796 (2007).

[456] Ramsey, S., Orrell, D. and Bolouri, H. *Dizzy: stochastic simulation of large-scale genetic regulatory networks.* J. Bioinform. Comput. Biol. **3**:415–436 (2005).

[457] Rao, C. V., Sauro, H. M. and Arkin, A. P. *Putting the "Control" in Metabolic Control Analysis.* DYCOPS (2004).

[458] Rao, C. V., Wolf, D. M. and Arkin, A. P. *Control, exploitation and tolerance of intracellular noise.* Nature **420**:231–237 (2002).

[459] Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmuller, U. and Timmer, J. *Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood.* Bioinformatics **25**:1923–1929 (2009).

[460] Ravasz, E. and Barabási, A.-L. *Hierarchical organization in complex networks.* Physical Review E **67**:026112 (2003).

[461] Ravasz, E., Somera, A., Mongru, D., Oltvai, Z. and Barabási, A.-L. *Hierarchical Organization of Modularity in Metabolic Networks.* Science **297**:1551–1555 (2002).

[462] Reddy, V. N., Mavrovouniotis, M. L. and Liebman, M. N. *Petri net representations in metabolic pathways.* Proc. Int. Conf. Intell. Syst. Mol. Biol. **1**:328–336 (1993).

[463] Reed, W. J. and Hughes, B. D. *From gene families and genera to incomes and internet file sizes: why power laws are so common in nature.* Phys. Rev. E **66**:067103 (2002).

[464] Reguly, T., Breitkreitz, A., Boucher, L., Breitkreutz, B.-J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskya, O. G., Ideker, T., Dolinski, K., Batada, N. N. and Tyers, M. *Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae.* Journal of Biology **5**:11 (2006).

[465] Reichardt, J. and Bornholdt, S. *Detecting fuzzy community structures in complex networks with a Potts model.* Phys. Rev. Lett. **93**:218701 (2004).

[466] Reichardt, J. and Bornholdt, S. *Statistical mechanics of community detection.* Phys. Rev. E **74**:016110 (2006).

[467] Reinitz, J., Mjolsness, E. and Sharp, D. H. *Model for cooperative control of positional information in* Drosophila *by bicoid and maternal hunchback.* J. Exp. Zool. **271**:47–56 (1995).

[468] Reinitz, J. and Sharp, D. H. *Mechanism of eve stripe formation.* Mech. Dev. **49**:133–158 (1995).

[469] Reinker, S., Altman, R. M. and Timmer, J. *Parameter estimation in stochastic biochemical reactions.* Syst. Biol. **153**:168–178 (2006).

[470] Ribeiro, A., Zhu, R. and Kauffman, S. A. *A general modeling strategy for gene regulatory networks with stochastic dynamics.* J. Comput. Biol. **13**:1630–1639 (2006).

[471] Richardson, T., Mucha, P. J. and Porter, M. A. *Spectral tripartitioning of networks.* Phys. Rev. E **80**:036111 (2009).

[472] Rives, A. W. and Galitski, T. *Modular organization of cellular networks.* PNAS **100**:1128–1133 (2003).

[473] Rodriguez-Fernandez, M. and Banga, J. R. *SensSB: A software toolbox for the development and sensitivity analysis of systems biology models.* Bioinformatics (2010).

[474] Rodriguez-Fernandez, M., Egea, J. A. and Banga, J. R. *Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems.* BMC Bioinformatics **7**:483 (2006).

[475] Rodriguez-Fernandez, M., Mendes, P. and Banga, J. R. *A hybrid approach for efficient and robust parameter estimation in biochemical pathways.* Biosystems **83**:248–65 (2006).

[476] Rohr, C., Marwan, W. and Heiner, M. *Snoopy–a unifying Petri net framework to investigate biomolecular networks.* Bioinformatics **26**:974–975 (2010).

[477] Roper, R. T., Pia Saccomani, M. and Vicini, P. *Cellular signaling identifiability analysis: a case study.* J. Theor. Biol. **264**:528–537 (2010).

[478] Ropers, D., de Jong, H., Page, M., Schneider, D. and Geiselmann, J. *Qualitative simulation of the carbon starvation response in Escherichia coli.* BioSystems **84**:124–152 (2006).

[479] Rosenfeld, N. and Alon, U. *Response delays and the structure of transcription networks.* J. Mol. Biol. **329**:645–654 (2003).

[480] Rosenfeld, N., Elowitz, M. B. and Alon, U. *Negative autoregulation speeds the response times of transcription networks.* J. Mol. Biol. **323**:785–793 (2002).

[481] Rosvall, M. and Bergstrom, C. T. *An information-theoretic framework for resolving community structure in complex networks.* PNAS **104**:7327–7331 (2007).

[482] Roussel, M. R. and Zhu, R. *Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression.* Phys. Biol. **3**:274–284 (2006).

[483] Ruan, J. and Zhang, W. *Identifying network communities with a high resolution.* Phys. Rev. E **77**:016104 (2008).

[484] Ruths, D., Muller, M., Tseng, J. T., Nakhleh, L. and Ram, P. T. *The signaling petri net-based simulator: a non-parametric strategy for characterizing the dynamics of cell-specific signaling networks.* PLoS Comput. Biol. **4**:e1000005 (2008).

[485] Saarinen, A., Linne, M. L. and Yli-Harja, O. *Stochastic differential equation model for cerebellar granule cell excitability.* PLoS Comput. Biol. **4**:e1000004 (2008).

[486] Sabidussi, G. *The centrality of a graph.* Psychometrika **31**:581–603 (1966).

[487] Saccomani, M. P., Audoly, S., Bellu, G. and D'Angio, L. *Examples of testing global identifiability of biological and biomedical models with the DAISY software.* Comput. Biol. Med. **40**:402–407 (2010).

[488] Sackmann, A., Formanowicz, D., Formanowicz, P. and Blazewicz, J. *New insights into the human body iron metabolism analyzed by a Petri net based approach.* BioSystems **96**:104–113 (2009).

[489] Saez-Rodriguez, J., Goldsipe, A., Muhlich, J., Alexopoulos, L. G., Millard, B., Lauffenburger, D. A. and Sorger, P. K. *Flexible informatics for linking experimental data to mathematical models via* `DataRail`. Bioinformatics **24**:840–847 (2008).

[490] Saez-Rodriguez, J., Hammerle-Fickinger, A., Dalal, O., Klamt, S., Gilles, E. D. and Conradi, C. *Multistability of signal transduction motifs.* IET Syst. Biol. **2**:80–93 (2008).

[491] Sahle, S., Mendes, P., Hoops, S. and Kummer, U. *A new strategy for assessing sensitivities in biochemical models.* Philos. Transact. A Math. Phys. Eng. Sci. **366**:3619–3631 (2008).

[492] Salathé, M., May, R. M. and Bonhoeffer, S. *The evolution of network topology by selective removal.* J. R. Soc. Interface **2**:533–536 (2005).

[493] Sales-Pardo, M., Guimerà, R., Moreira, A. A. and Amaral, L. A. *Extracting the hierarchical organization of complex systems.* PNAS **104**:15224–15229 (2007).

[494] Salis, H., Sotiropoulos, V. and Kaznessis, Y. N. *Multiscale `Hy3S`: hybrid stochastic simulation for supercomputers.* BMC Bioinformatics **7**:93 (2006).

[495] Saltelli, A., Ratto, M., Tarantola, S. and Campolongo, F. *Sensitivity analysis for chemical models.* Chem. Rev. **105**:2811–2828 (2005).

[496] Saltelli, A., Tarantola, S. and Campolongo, F. *Sensitivity Anaysis as an Ingredient of Modeling.* Statist. Sci. **15**(4):377–395 (2000).

[497] Samad, H., Khammash, M., Petzold, L. and Gillespie. *Stochastic modelling of gene regulatory networks.* Int. J. Robust Nonlinear Control **15**:691–711 (2005).

[498] Santos, V. J., Bustamante, C. D. and Valero-Cuevas, F. J. *Improving the fitness of high-dimensional biomechanical models via data-driven stochastic exploration.* IEEE Trans. Biomed. Eng. **56**:552–564 (2009).

[499] Sauro, H. M., Hucka, M., Finney, A., Wellock, C., Bolouri, H., Doyle, J. and Kitano, H. *Next generation simulation tools: the Systems Biology Workbench and `BioSPICE` integration.* OMICS **7**:355–372 (2003).

[500] Savageau, M. A. *Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions.* J. Theor. Biol. **25**:365–369 (1969).

[501] Savageau, M. A. *Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation.* J. Theor. Biol. **25**:370–379 (1969).

[502] Savageau, M. A. *Biochemical systems analysis. 3. Dynamic solutions using a power-law approximation.* J. Theor. Biol. **26**:215–226 (1970).

[503] Savageau, M. *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology.* Addison-Wesley (1976).

[504] Savageau, M. *Introduction to S-systems and the underlying power-law formalism.* Math. Comput. Modelling **2**:546–551 (1988).

[505] Savageau, M. and Voit, E. *Power-Law Approach to Modeling Biological-Systems: I. Theory.* J. of Ferment. Technol. **60(3)**:221–228 (1982).

[506] Schaibly, J. and Shuler, K. *Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. Part II: applications.* Journal of Chemical Physics **59**(8):3879–3888 (1973).

[507] Scheper, T., Klinkenberg, D., Pennartz, C. and van Pelt, J. *A mathematical model for the intracellular circadian rhythm generator.* J. Neurosci. **19**:40–47 (1999).

[508] Schilstra, M. J., Li, L., Matthews, J., Finney, A., Hucka, M. and Le Novère, N. *`CellML2SBML`: conversion of CellML into SBML.* Bioinformatics **22**:1018–1020 (2006).

[509] Schmidt, H. and Jirstrand, M. *Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology.* Bioinformatics **22**:514–515 (2006).

[510] Schmidt, H., Madsen, M. F., Danø, S. and Cedersund, G. *Complexity reduction of biochemical rate expressions.* Bioinformatics **24**:848–854 (2008).

[511] Schulz, M., Uhlendorf, J., Klipp, E. and Liebermeister, W. *`SBMLmerge`, a system for combining biochemical network models.* Genome Inform. **17**:62–71 (2006).

[512] Schwartz, R. *Biological Modeling and Simulation.* MIT Press (2008).

[513] Seidmann, S. B. *Network structure and minimum degree.* Social Networks **5**:269–287 (1983).

[514] Selivanov, V. A., de Atauri, P., Centelles, J. J., Cadefau, J., Parra, J., Cussó, R., Carreras, J. and Cascante, M. *The changes in the energy metabolism of human muscle induced by training.* J. Theor. Biol. **252**:402–410 (2008).

[515] Sen, T. Z., Kloczkowski, A. and Jernigan, R. L. *Functional clustering of yeast proteins from the protein-protein interaction network.* BMC Bioinformatics **7**:355 (2006).

[516] Shen, J. *A direct method of calculating sensitivity coefficients of chemical kinetics.* J. Chem. Phys. **31**:64–68 (2002).

[517] Shen-Orr, S., Milo, R., Mangan, S. and Alon, U. *Network motifs in the transcriptional regulation network of* Escherichia coli. Nature Genetics **31**:64–68 (2002).

[518] Shinar, G. and Feinberg, M. *Structural sources of robustness in biochemical reaction networks.* Science **327**:1389–1391 (2010).

[519] Shmulevich, I. and Zhang, W. *Binary analysis and optimization-based normalization of gene expression data.* Bioinformatics **18**:555–565 (2002).

[520] Simon, H. A. *On a class of skew distribution functions.* Biometrika **42**(3/4):425–440 (1955).

[521] Smolen, P., Baxter, D. A. and Byrne, J. H. *A reduced model clarifies the role of feedback loops and time delays in the* Drosophila *circadian oscillator.* Biophys. J. **83**:2349–2359 (2002).

[522] Smolen, P., Hardin, P. E., Lo, B. S., Baxter, D. A. and Byrne, J. H. *Simulation of* Drosophila *circadian oscillations, mutations, and light responses by a model with VRI, PDP-1, and CLK.* Biophys. J. **86**:2786–2802 (2004).

[523] Snoussi, E. *Qualitative dynamics of piece-linear differential equations: a discrete mapping approach.* Dyn. Stability Systems **4**:189–207 (1989).

[524] Sobol, I. *Sensitivity analysis for non-linear mathematical models.* Mathematical Modeling & Computational Experiment **1**:407–414 (1993).

[525] Sobol, I. *Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates.* Mathematics and Computers in Simulation **55**(1-3):271–280 (2001).

[526] Soffer, S. N. and Vázquez, A. *Network clustering coefficient without degree-correlation biases.* Phys. Rev. E **71**:057101 (2005).

[527] Solé, R. V., Pastor-Satorras, R., Smith, E. D. and Kepler, T. *A model of large-scale proteome evolution.* Advances in Complex Systems **5**:43–54 (2002).

[528] Somogyi, R. and Sniegoski, C. *Modeling the complexity of genetic networks: Understanding multigenic and pleiotropic regulation.* Complexity **1**(6):45–63 (1996).

[529] Song, J. and Singh, M. *How and when should interactome-derived clusters be used to predict functional modules and protein function?* Bioinformatics **25**:3143–3150 (2009).

[530] Sorribas, A. and González, A. *The Power-law Formalism as a Tool for Modeling Hormonal Systems.* Journal of Theoretical Medicine **2**:19–38 (1999).

[531] Sorribas, A., Hernandez-Bermejo, B., Vilaprinyo, E. and Alves, R. *Cooperativity and saturation in biochemical networks: a saturable formalism using Taylor series approximations.* Biotechnol. Bioeng. **97**:1259–1277 (2007).

[532] Spirin, A. S., Serdyuk, I. N., Shpungin, J. L. and Vasiliev, V. D. *Quaternary structure of the ribosomal 30S subunit: model and its experimental testing.* PNAS **76**:4867–4871 (1979).

[533] Spirin, V. and Mirny, L. A. *Protein complexes and functional modules in molecular networks.* PNAS **100**:12123–12128 (2003).

[534] Srinath, S. and Gunawan, R. *Parameter identifiability of power-law biochemical system models.* J. Biotechnol. (2010).

[535] Srinivas, M. and Patnaik, L. M. *Genetic Algorithms: A Survey.* IEEE COMPUTER pages 17–26 (1994).

[536] Stanley, M. H. R., Amaral, L. A. N., Buldyrev, S. V., Havlin, S., Leschhorn, H., Maass, P., Salinger, M. A. and Stanley, H. *Scaling behaviour in the growth of companies.* Nature **379**:804–806 (1996).

[537] Steggles, L. J., Banks, R., Shaw, O. and Wipat, A. *Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach.* Bioinformatics **23**:336–343 (2007).

[538] Stelling, J., Gilles, E. D. and Doyle III, F. J. *Robustness properties of circadian clock architectures.* PNAS **101**:13210–13215 (2004).

[539] Stelling, J., Sauer, U., Szallasi, Z., Doyle III, F. J. and Doyle, J. *Robustness of Cellular Functions.* Cell **118**:675–685 (2004).

[540] Stephenson, K. A. and Zelen, M. *Rethinking centrality: Methods and examples.* Social Networks **11**:1–37 (1989).

[541] Strogatz, S. H. *Nonlinear Dynamics and Chaos.* Addison Wesley (1994).

[542] Strogatz, S. H. *Exploring complex networks.* Nature **410**:268–276 (2001).

[543] Stumpf, M. P., Kelly, W. P., Thorne, T. and Wiuf, C. *Evolution at the system level: the natural history of protein interaction networks.* Trends Ecol. Evol. **22(7)**:366–373 (2007).

[544] Stumpf, M. P. H. and Ingram, P. *Probability Models for Degree Distributions of Protein Interaction Networks.* Europhys. Lett. **71**:152158 (2005).

[545] Stumpf, M. P. H., Wiuf, C. and May, R. M. *Subnets of scale-free networks are not scale-free: Sampling properties of networks.* PNAS **102**:4221–4224 (2005).

[546] Sugimoto, M., Kikuchi, S. and Tomita, M. *Reverse engineering of biochemical equations from time-course data by means of genetic programming.* BioSystems **80**:155–164 (2005).

[547] Sun, X., Jin, L. and Xiong, M. *Extended kalman filter for estimation of parameters in nonlinear state-space models of biochemical networks.* PLoS ONE **3**:e3758 (2008).

[548] Swameye, I., Muller, T. G., Timmer, J., Sandra, O. and Klingmuller, U. *Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling.* PNAS **100**:1028–1033 (2003).

[549] Szczurek, E., Gat-Viks, I., Tiuryn, J. and Vingron, M. *Elucidating regulatory mechanisms downstream of a signaling pathway using informative experiments.* Mol. Syst. Biol. **5**:287 (2009).

[550] Tanaka, R., Yi, T.-M. and Doyle, J. *Some protein interaction data do not exhibit power law statistics.* FEBS Letters **579**:5140–5144 (2005).

[551] Taubes, C. *Modeling Differential Equations in Biology.* Prentice Hall (2000).

[552] Thieffry, D. and Thomas, R. *Qualitative analysis of gene networks.* Pac Symp Biocomput pages 77–88 (1998).

[553] Thomas, A. *Regulatory networks seen as asynchronous automata: A logical description.* J. Theor. Biol. **153**(1):1–23 (1991).

[554] Thomas, A., Cannings, R., Monk, N. A. and Cannings, C. *On the structure of protein-protein interaction networks.* Biochem. Soc. Trans. **31**:1491–1496 (2003).

[555] Thomas, A. and D'Ari, R. *Biological Feedback.* CRC Press (1990).

[556] Tian, T., Burrage, K., Burrage, P. and Carletti, M. *Stochastic delay differential equations for genetic regulatory networks.* Journal of Computational and Applied Mathematics **205**:696–707 (2007).

[557] Tomshine, J. and Kaznessis, Y. N. *Optimization of a stochastically simulated gene network model via simulated annealing.* Biophys. J. **91**:3196–3205 (2006).

[558] Tong, W. C. and Holden, A. V. *Bifurcation analysis of genetically engineered pacemaking in Mammalian heart.* J. Biol. Phys. **32**:169–172 (2006).

[559] Tonnelier, A. *The McKean's caricature of the FitzHugh-Nagumo model. The space-clamped system.* SIAM J. Appl. Math. **63**:459–484 (2002).

[560] Tonnelier, A. *McKean caricature of the FitzHugh-Nagumo model: traveling pulses in a discrete diffusive medium.* Phys. Rev. E **67**:036105 (2003).

[561] Tonnelier, A. and Gerstner, W. *Piecewise linear differential equations and integrate-and-fire neurons: insights from two-dimensional membrane models.* Phys. Rev. E **67**:021908 (2003).

[562] Torres, N. and Voit, E. *Pathway Analysis and Optimization in Metabolic Engineering.* Cambridge University Press (2003).

[563] Tsai, K. Y. and Wang, F. S. *Evolutionary optimization with data collocation for reverse engineering of biological networks.* Bioinformatics **21**:1180–1188 (2005).

[564] Turányi, T. *Sensitivity analysis of complex kinetic systems. Tools and applications.* Journal of Mathematical Chemistry **5**:203–248 (1990).

[565] Turing, A. *The chemical basis of morphogenesis.* Phil. Trans. Roy. Soc. B **237**:3772 (1952).

[566] Tyson, J. J., Chen, K. and Novák, B. *Network dynamics and cell physiology.* Nat. Rev. Mol. Cell Biol. **2**:908–916 (2001).

[567] Tyson, J. J., Chen, K. C. and Novák, B. *Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell.* Curr. Opin. Cell Biol. **15**:221–231 (2003).

[568] Ueda, H. R., Hayashi, S., Matsuyama, S., Yomo, T., Hashimoto, S., Kay, S. A., Hogenesch, J. B. and Iino, M. *Universality and flexibility in gene expression from bacteria to human.* PNAS **101**:3765–3769 (2004).

[569] van Someren, E. P., Wessels, L. F. and Reinders, M. J. *Linear modeling of genetic networks from experimental data.* Proc. Int. Conf. Intell. Syst. Mol. Biol. **8**:355–366 (2000).

[570] Vázquez, A. *Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations.* Phys. Rev. E **67**:056104 (2003).

[571] Vázquez, A., Flammini, A., Maritan, A. and Vespignani, A. *Modeling of Protein Interaction Networks.* ComPlexUs **1**:38–44 (2003).

[572] Vera, J., Balsa-Canto, E., Wellstead, P., Banga, J. R. and Wolkenhauer, O. *Power-law models of signal transduction pathways.* Cell. Signal. **19**:1531–1541 (2007).

[573] Versari, C. and Busi, N. *Stochastic biological modelling in the presence of multiple compartments.* Theoretical Computer Science **410**:3039–3064 (2009).

[574] Vilela, M., Borges, C. C., Vinga, S., Vasconcelos, A. T., Santos, H., Voit, E. O. and Almeida, J. S. *Automated smoother for the numerical decoupling of dynamics models.* BMC Bioinformatics **8**:305 (2007).

[575] Vilela, M., Chou, I. C., Vinga, S., Vasconcelos, A. T., Voit, E. O. and Almeida, J. S. *Parameter optimization in S-system models.* BMC Syst. Biol. **2**:35 (2008).

[576] Voit, E. O. and Almeida, J. *Decoupling dynamical systems for pathway identification from metabolic profiles.* Bioinformatics **20**:1670–1681 (2004).

[577] Voit, E. *Canonical non-linear modeling: S-system approach to understanding complexity.* Van Nostrand Reinhold (1991).

[578] Voit, E. *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists.* Cambridge University Press (2000).

[579] Voit, E. and Savageau, M. *Power-Law Approach to Modeling Biological-Systems: II. Application to Ethanol-Production.* J. of Ferment. Technol. **60**(3):229–232 (1982).

[580] Volterra, V. *Variazioni e fluttuazioni del numero dindividui in specie animali conviventi.* Mem. Acad. Lineci Roma **2**:31–113 (1926).

[581] Volterra, V. *Variations and fluctuations of the number of individuals in animal species living together in Animal Ecology.* Chapman, R.N., McGrawHill (1931).

[582] von Dassow, G., Meir, E., Munro, E. M. and Odell, G. M. *The segment polarity network is a robust developmental module.* Nature **406**:188–192 (2000).

[583] Voss, K., Heiner, M. and Koch, I. *Steady state analysis of metabolic pathways using Petri nets.* In Silico Biol. **3**:367–387 (2003).

[584] Wagner, A. *Robustness against mutations in genetic networks of yeast.* Nat. Genet. **24**:355–361 (2000).

[585] Wagner, A. *How the global structure of protein interaction networks evolves.* Proc. Biol. Sci. **270**:457–466 (2003).

[586] Wagner, A. *Robustness and Evolvability in Living Systems.* Princeton University Press (2005).

[587] Wagner, A. and Fell, D. A. *The small world inside large metabolic networks.* Proc. Biol. Sci. **268**:1803–1810 (2001).

[588] Wagner, J., Ma, L., Rice, J. J., Hu, W., Levine, A. J. and Stolovitzky, G. A. *p53-Mdm2 loop controlled by a balance of its feedback strength and effective dampening using ATM and delayed feedback.* Syst. Biol. **152**:109–118 (2005).

[589] Wahde, M. and Hertz, J. *Coarse-grained reverse engineering of genetic regulatory networks.* BioSystems **55**:129–136 (2000).

[590] Walter, E. and Pronzato, L. *Identification of Parametric Models from Experimental Data.* Springer (1997).

[591] Wasserman, S. and Faust, K. *Social Network Analysis: Methods and Applications.* Cambridge University Press (1994).

[592] Watts, D. J. and Strogatz, S. H. *Collective dynamics of "small-world" networks.* Nature **393**:440–442 (1998).

[593] Weaver, D. C., Workman, C. T. and Stormo, G. D. *Modeling regulatory networks with weight matrices.* Pac. Symp. Biocomput. pages 112–123 (1999).

[594] Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. and Somogyi, R. *Large-scale temporal gene expression mapping of central nervous system development.* PNAS **95**:334–339 (1998).

[595] Wessels, L. F., van Someren, E. P. and Reinders, M. J. *A comparison of genetic network models.* Pac. Symp. Biocomput. pages 508–519 (2001).

[596] Wildermuth, M. C. *Metabolic control analysis: biological applications and insights.* Genome Biol. **1**(6):1031.1–1031.5 (2000).

[597] Wilkinson, D. J. *Stochastic modelling for quantitative description of heterogeneous biological systems.* Nat. Rev. Genet. **10**:122–133 (2009).

[598] Wilkinson, D. M. and Huberman, B. A. *A method for finding communities of related genes.* PNAS **101**(Suppl. 1):5241–5248 (2004).

[599] Wilkinson, D. *Stochastic Modelling for Systems Biology.* Chapman & Hall/CRC (2006).

[600] Wolfram, S. *A New Kind of Science.* Wolfram Media (2002).

[601] Wu, F. and Huberman, B. A. *Finding communities in linear time: a physics approach.* The European Physical Journal B **38**(2):331–338 (2004).

[602] Wu, J., Tan, Y., Deng, H. and Zhu, D. *Relationship between degree-rank function and degree distribution of protein-protein interaction networks.* Comput. Biol. Chem. **32**:1–4 (2008).

[603] Wu, J. and Voit, E. *Hybrid modeling in biochemical systems theory by means of functional petri nets.* J. Bioinform. Comput. Biol. **7**:107–134 (2009).

[604] Wu, W. H., Wang, F. S. and Chang, M. S. *Dynamic sensitivity analysis of biological systems.* BMC Bioinformatics **9**(Suppl. 12):S17 (2008).

[605] Wuchty, S. and Almaas, E. *Evolutionary cores of domain co-occurrence networks.* BMC Evol. Biol. **5**:24 (2005).

[606] Wuchty, S. and Almaas, E. *Peeling the yeast protein network.* Proteomics **5**:444–449 (2005).

[607] Wuchty, S. and Stadler, P. F. *Centers of complex networks.* J. Theor. Biol. **223**:45–53 (2003).

[608] Wuensche, A. *Genomic regulation modeled as a network with basins of attraction.* Pac. Symp. Biocomput. pages 89–102 (1998).

[609] Yang, X.-S. *Computational modelling of nonlinear calcium waves.* Applied Mathematical Modelling **30**:200–208 (2006).

[610] Yoon, J. and Deisboeck, T. S. *Investigating differential dynamics of the MAPK signaling cascade using a multi-parametric global sensitivity analysis.* PLoS ONE **4**:e4560 (2009).

[611] Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X. and Gerstein, M. *Genomic analysis of essentiality within protein networks.* Trends Genet. **20**:227–231 (2004).

[612] Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. and Gerstein, M. *The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.* PLoS Comput. Biol. **3**:e59 (2007).

[613] Yue, H., Brown, M., He, J., F. Jia and Kell, D. *Sensitivity Analysis and Robust Experimental Design of a Signal Transduction Pathway System.* International Journal of Chemical Kinetics **40**:730–741 (2008).

[614] Yue, H., Brown, M., Knowles, J., Wang, H., Broomhead, D. and Kell, D. *Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: a case study of an NF-κB signalling pathway.* Mol Biosyst **2**:640–649 (2006).

[615] Yuh, C. H., Bolouri, H. and Davidson, E. H. *Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene.* Science **279**:1896–1902 (1998).

[616] Yuh, C. H., Moore, J. G. and Davidson, E. H. *Quantitative functional interrelations within the cis-regulatory system of the S. purpuratus Endo16 gene.* Development **122**:4045–4056 (1996).

[617] Zager, M. G., Schlosser, P. M. and Tran, H. T. *A delayed nonlinear PBPK model for genistein dosimetry in rats.* Bull. Math. Biol. **69**:93–117 (2007).

[618] Zak, D. E., Gonye, G. E., Schwaber, J. S. and Doyle, F. J. *Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network.* Genome Res. **13**:2396–2405 (2003).

[619] Zhang, T., Zhang, D., Li, Z. and Cai, Q. *Evaluating the structural identifiability of the parameters of the EBPR sub-model in ASM2d by the differential algebra method.* Water Res. **44**:2815–2822 (2010).

[620] Zheng, Y. and Rundell, A. *Comparative study of parameter sensitivity analyses of the TCR-activated Erk-MAPK signalling pathway.* Syst. Biol. **153**:201–211 (2006).

[621] Zhu, R., Ribeiro, A. S., Salahub, D. and Kauffman, S. A. *Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models.* J. Theor. Biol. **246**:725–745 (2007).

[622] Zhu, X., Gerstein, M. and Snyder, M. *Getting connected: analysis and principles of biological networks.* Genes & Dev. **21**:1010–1024 (2007).

[623] Zi, Z., Cho, K. H., Sung, M. H., Xia, X., Zheng, J. and Sun, Z. *In silico identification of the key components and steps in IFN-γ induced JAK-STAT signaling pathway.* FEBS Lett. **579**:1101–1108 (2005).

[624] Zi, Z. and Klipp, E. *SBML-PET: A Systems Biology Markup Language based Parameter Estimation Tool.* Bioinformatics **22**(21):2704–2705 (2006).

[625] Zi, Z. and Sun, Z. *Robustness Analysis of the IFN-γ Induced JAK-STAT Signaling Pathway.* J. Comput. Sci. & Technol. **20**(4):491–495 (2005).

[626] Zi, Z., Zheng, Y., Rundell, A. E. and Klipp, E. *SBML-SAT: a systems biology markup language (SBML) based sensitivity analysis tool.* BMC Bioinformatics **9**:342 (2008).

[627] Zotenko, E., Mestre, J., O'Leary, D. P. and Przytycka, T. M. *Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality.* PLoS Comput. Biol. **4**:e1000140 (2008).

[628] Zwolak, J. W., Tyson, J. J. and Watson, L. T. *Globally optimised parameters for a model of mitotic control in frog egg extracts.* Syst. Biol. **152**:81–92 (2005).

Glossary

Adjacency matrix

A real symmetric matrix A used as representation of an undirected graph $G = (\mathcal{V}, \mathcal{E})$. The elements of A are $a_{i,j}$ taking the values 1 if there is an edge between vertex i and j, 0 otherwise. $\{i, j\} \in \mathcal{V}$. 14, 19

Artificial neural network

Computational modelling technique consisting on simulating a system using a network of nodes which behaviour mimics to a certain point the functions of real neurons. A node works as the input of the system and the rest of nodes as functions $f : X \rightarrow Y$. The output of the system is a composition of the other functions, $f(x) = K\Big(\sum_i \omega_i g_i(x)\Big)$. Some of the most popular neural networks are support vector machines and self-organising maps. The amount of applications is broad, specially efficient on learning problems as classification, pattern recognition or image processing. 27

Bayesian Network

A Bayesian network is a probabilistic graph model $G = (\mathcal{V}, \mathcal{E})$, being G an acyclic graph where vertices $v \in \mathcal{V}$ are represented by random variables x_i and edges $\mathcal{E}$ are conditional dependencies. They have been applied to network inference and learning, specifically in computational biology, where their use is widely common in network inference from microarray experiments [187]. 27

Bifurcation analysis

Bifurcation analysis consists on a mathematical analysis of topologically inequivalent changes on dynamical systems. 32, 37

BioModels Database

Data repository of mathematical models of biochemical interest [322]. Models are stored in SBML format. Models are annotated and linked to other relevant resources as publications, ontologies and other biochemical databases. Models can be browsed, downloaded, exported into other formats, visualised and simulated. Web services are also provided. 28, 29, 54, 60, 61, 144

Boolean Network

Proposed by S. A. Kauffman [292], a Boolean network is a discrete dynamical modelling method. States and time are defined as discrete. Cellular automata is a particular case of Boolean networks. As in the original work, it has been further applied to gene regulatory networks [528, 608, 267, 519]. Each of the genes are represented by a system of N-binary variables with K inputs as the regulatory mechanisms, being 2^N possible states, arriving eventually to an atractor. Different variable update mechanisms have been defined. 27

ByoDyn

Command line program for the analysis of dynamical systems of biochemical interest. It inputs as primary model format SBML models and the software holds a large variety of analysis tools. Some of the most relevant ones are deterministic and stochastic simulations, sensitivity analysis, model calibration, identifiability analysis and optimal experimental design. Some of the most computationally expensive routines have been parallelised, specifically some of the global methods for model calibration. The program is mainly written in Python and its sources are freely available for the scientific community. It is distributed from SourceForge under the GNU GPL Licence. It has been developed, tested and used under Mac OS and Linux operating systems. 56, 58–61, 94, 95

ByoDynWeb

Graphical interface for `ByoDyn`. It is a web-based server that allows the user to send a calculation of `ByoDyn` using a graphical interface. The user needs first to register (for free) and they will access a personal workspace where different models and runners are available. All runners from the `ByoDyn` documentation are available and a large amount of models from public repositories (BioModels Database and JWS Online Model Database) are also directly uploaded into the server, accessible for the user. The calculations can be sent using the displayed windows to several computational environments: (1) the web server itself providing robustness under network failures, (2) a dedicated PC for regular calculations and (3) a dedicated cluster of machines for specially long calculations. 61

CellML
: It is an open standard computational language based on the XML markup language. It is intended to store and exchange computer-based mathematical models, specifically for different aspects of biology [345]. Tools and model repositories are publicly available. 28, 53

Cellular automanton
: Cellular automata is a discrete modelling technique. It consists on a regular grid of elements called cells with specified rules of connectivity and evolution. As time (also discrete) advances, the state of a cell variates depending on the state of its neighbours. Generally update rules are deterministic although probabilistic ones are also possible. Cell's states are in principle finite but continuous automata have been described too. 27

Clique
: A clique, in an undirected graph, is any of their complete subgraphs, i.e., a subset of vertices so that every two vertices are connected. 13, 14

Clustering coefficient
: Local property of a vertex from a graph. It provides an idea of the local cohesiveness of the vertex evaluating the number of edges among the direct neighbours of the given vertex. It can also be called *transitivity*. 5

Degree distribution
: Global property of a graph, generally noted as $P(k)$, consisting on the probability distribution of the degree over all vertices of the graph. The degree k of a vertex is no more than the number of edges with the other vertices of the graph. 4, 148

Determinant of a matrix
: Defined on a n-by-n square matrix, it is a mathematical object defined as,

$$\det(A) = \sum_{j=1}^{n} A_{i,j}(-1)^{i+j} M_{i,j} \tag{6.1}$$

being $M_{i,j}$ the minor, which is defined as the determinant of the matrix resulting from removing the i-row and the j-column of A.

Geometrically it can be interpreted as a scale factor for measure. 50

Difference Equations
Not to be confused with differential equations, difference equations are a specific type of recurrence relation, where an equation defines a sequence of recurring terms, $x_i = f(t, x_{i-1})$. It has been applied to modelling dynamical systems in discrete timescales [368, 552]. 28

Differential Equations
A differential equation is an equation involving the unknown functions and its derivatives.

$$y' = f(x, y) \tag{6.2}$$

Several types of differential equations are explained in Section 1.3, including ordinary differential equations, linear and non-linear differential equations, partial differential equations, algebraic differential equations, delay differential equations, piecewise differential equations, stochastic differential equations and complex differential equations. 146

Edge
As defined in [407], an edge is the line connecting two vertices. In a social network it represents a relationship, a wire in a computer network or a functional relationship in a transcriptional regulatory network. 2

Fortran
Fortran is a programming language developed by IBM in the 1950s. It is specially suitable for numerical computation. 53, 56

Fourier amplitude sensitivity test
It is a sensitivity method based on the Fourier transformation reducing the multidimensional input model into a single dimensional one. From a sample of inputs, the expected value and the expected variance (along with the contribution of each parameter to it) are calculated. 37

Gaussian-Newton method

The Gaussian-Newton method is a numerical algorithm to solve non-linear least squares problems. Given a model,

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\theta}), \tag{6.3}$$

and the associated sum of squares,

$$SS(\boldsymbol{\theta}) = \sum_{i=1}^{l} r_i^2(\boldsymbol{\theta}). \tag{6.4}$$

The new position at the parameter space for iteration $t+1$ is taken as,

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^t + \boldsymbol{\delta}, \tag{6.5}$$

where $\boldsymbol{\delta}$ can be taken from,

$$(\boldsymbol{J}^T\boldsymbol{J})\boldsymbol{\delta} = \boldsymbol{J}^T\boldsymbol{r}, \tag{6.6}$$

being $\boldsymbol{J}^T$ the transpose of the Jacobian matrix of f with respect to $\boldsymbol{\theta}$. 43, 149

Genetic algorithm

Genetic algorithms are a type of global search algorithms inspired by natural selection. The idea is the following: a group of elements are initialised typically at random positions of the parameter space. Each of the elements are evaluated and a fitness value is associated with them. Those elements with better fitness pass to the next generation. The selected elements can undergo modification procedures with specific probabilities. Those modifications are called mutation and cross-over in which an element is modified or two elements are combined, respectively. Introduced and formalised by many scientists in the 60s and 70s a great variety of different modifications have been implemented [260, 217, 122, 388]. In the field of computational systems biology, they have been applied extensively [535, 426, 589, 298, 546]. 44, 52

Geometric random graphs

A geometric random graph is a graph $G(\mathcal{V}, \delta)$, being δ a threshold distance. $\mathcal{V}$ is a set of vertices embedded in a metric space and $\mathcal{E}$ a set of edges such that $\mathcal{E} = \{(u, v) \in \mathcal{E} \mid \{u, v\} \in \mathcal{V} \wedge 0 < ||u - v|| \leq$

$\delta\}$, where $||\cdot||$ is an arbitrary distance norm in the metric space [447].

In other words, geometric random graphs are a type of random graphs constructed as follows:

1. In a bounded N-dimensional space a number of points are placed uniformly and independently.
2. Each point is referred as a graph vertex and two vertices are connected if the distance between them is below a given threshold.

Further information can be found at [431]. 19, 24

GNU GPL Licence

GNU General Public Licence is a software licence for free software. 144

Graphlet degree distribution

The graphlet degree distribution of a graph is the degree distribution of a graphlet. A graphlet is defined as a small connected non-isomorphic subgraph. See [446] for further insights. 24

High dimensional model representation

High dimensional model representation is a finite expansion of a multivariate function,

$$f(x_1, x_2, ..., x_k) = f_0 + \sum_i f_i + \sum_i \sum_{j>i} f_{ij} + ... + f_{12...k}. \quad (6.7)$$

It is used for the calculation of global sensitivities by the Sobol's method. 37

Homoscedasticity

Two or more variables are said to be homoscedastic if they have the same finite variance. 39

Hub

A vertex with a relatively large number of edges, ideally $|\mathcal{E}|$. 18

JWS Online Model Database

Data repository of mathematical models of biochemical interest

[422]. Models are stored in SBML format. Models can be run on-line or downloaded. 54, 144

Laplacian matrix

A $|\mathcal{V}| \times |\mathcal{V}|$ matrix L, which diagonal values are the degree of node i and L_{ij} values take -1 if an edge exist between vertices i and j and 0 otherwise. 14

Latin hypercube sampling

Latin hypercube sampling (LHS) is a statistical technique generalising the concept of Latin squares for multiple dimensions first described by McKay in 1979 [371]. A Latin square consist on a n-square matrix in which n different elements appear exactly once for each row and column. LHS guarantees that the sampled set is more representative of the real variability than a pure random sample. 37, 49

Lattice

A lattice is a discrete subgroup which spans a vector space, generally in $\mathbb{R}^n$. It can be defined as,

$$\mathcal{L} = \left\{ \sum_{i=1}^{n} a_i v_i \mid a_i \in \mathbb{Z} \right\}, \tag{6.8}$$

where $\{v_1, ..., v_n\}$ is a basis for the vector space. 21

Leave-one-out cross validation

A statistical technique to evaluate the performance of a predictive model, assessing how the results of a statistical analysis will generalise to an independent data set. The idea is to divide a complete data set into complementary subsets, the training set and the validation set. The training set is used for analysis (parametrization for example) and the validation set is used for evaluation of the performance of the model. Commonly different partitions are created and tested, then an average result is adopted. 88

Levenberg-Marquardt method

The Levenberg-Marquardt method is a modification of the Gaussian-Newton in which the increment $\boldsymbol{\delta}$ can be solved from,

$$\big((\boldsymbol{J}^T\boldsymbol{J}) + \lambda \mathrm{diag}(\boldsymbol{J}^T\boldsymbol{J})\big)\boldsymbol{\delta} = \boldsymbol{J}^T\boldsymbol{r}, \tag{6.9}$$

where λ is a damping adjustable factor and $\text{diag}(\boldsymbol{J}^T\boldsymbol{J})$ is a scaling term that accomodates the curvature of the gradient. 43

Linux
It refers to the Unix-like operating systems based on the Linux kernel developed originally by Linus Torvalds in 1991. The underlying code is subjected to the GNU GPL Licence. It is now widespread on desktops, laptops and many other electronic devices. 144

Mac OS
Graphical operating system of Apple computers. 144

Maximum likelihood estimation
Given a sample data set of n independent observations $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, a maximum likelihood estimation is a method to determine which is the parameter value $\hat{\boldsymbol{\theta}}$ of a given model, more likely to produce such data. First a joint density function is constructed,

$$f(x_1, x_2, ..., x_n|\boldsymbol{\theta}) = f(x_1|\boldsymbol{\theta})f(x_2|\boldsymbol{\theta})...f(x_n|\boldsymbol{\theta}). \tag{6.10}$$

Reversing the roles, we arrive to the likelihood function,

$$l(\boldsymbol{\theta}|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}), \tag{6.11}$$

although more practically is to work with the log likelihood function,

$$L(\boldsymbol{\theta}|x_1, x_2, ..., x_n) = \ln l = \sum_{i=1}^{n} f(x_i|\boldsymbol{\theta}). \tag{6.12}$$

The parameter value $\hat{\boldsymbol{\theta}}$, that maximises the likelihood as,

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\ L(\boldsymbol{\theta}|x_1, x_2, ..., x_n) \tag{6.13}$$

is the most likely model parameter that explain the data set $\boldsymbol{x}$. Generally,

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 0 \tag{6.14}$$

is solved to find $\hat{\boldsymbol{\theta}}$. 12, 15, 40

Message passing interface (MPI)
It is a computational standard that specified the way of communicating among different computers. It is language-independent and it is the most used protocol for parallel applications in supercomputers and grids. 94

Metabolic network (MN)
Ensemble of metabolites related by the reaction they participate in. Generally a MN refers to the complete metabolism of an entire organism. 2

Metaheuristic
Computational algorithm that iteratively attempts to improve a measure quality based on simply defined rules. 44

Monte Carlo sampling
The Monte Carlo method refers to any method that starts from an ensemble of random points and select some of them based on some principles in order to achieve a given goal. 10, 12, 51

Motif
A network motif is a subgraph that occurs within a network more often than expected at random. 18, 19

Multifractal
A multifractal system refers to the generalisation of a fractal system, where more than a single exponent is required to describe it. 24

Partial rank correlation coefficient
A correlation coefficient consists on a measure of the degree of association between two variables. Given a sample of n elements for which two variables x and y are known, the correlation coefficient can be defined as,

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{6.15}$$

being $\bar{x}$ and $\bar{y}$ the mean values of x and y respectively. The partial correlation coefficient measures the correlation of a pair of variables (x_j, y) holding constant the influence of a third (x_k),

$$r_{x_j,y \bullet x_k} = \frac{r_{x_j,x_k} - r_{x_j,y} r_{x_k,y}}{\sqrt{(1 - r^2_{x_j,y})(1 - r^2_{x_k,y})}}. \tag{6.16}$$

The partial rank correlation coefficient is the result of a partial correlation on rank-transformed data. 37

Petri net

A Petri net is a mathematical formalism to model distributed systems. Formally is defined as a tuple $P = (\mathcal{S}, \mathcal{T}, \mathcal{W})$ where $\mathcal{S}$ is a set of places, $\mathcal{T}$ a set of transitions and $\mathcal{W}$ a set of archs. No object can be both a place and a transition. Arcs, which joint elements of $\mathcal{S}$ and $\mathcal{T}$ are defined as non-negative integer values. Introductory references are [434, 435]. 28

Phase transition

Generally a phase transition is the transformation of a thermodynamic system from one state of matter (where physical properties are essentially uniform) to another, e.g. from liquid to gas. Specifically for random graphs, it refers to the phenomenon that many monotone-increasing graph properties switch from very unlikely to highly likely around a threshold concerning the number of edges [158, 277]. 20

Pleiotropy

Pleiotropy is a genetic phenomenon where a single gene affects multiple phenotypic traits. A common example is the *CFTR* gene, whose mutations can cause multiple diseases, e.g., congenital absence of vas deferens or any of the many manifestations of cystic fibrosis: lung illness, gastrointestinal, liver or pancreatic failures, endocrine disorders, etc. 13

Power-law distribution

A power-law distribution is a probability distribution that satisfies the form:

$$p(x) \propto L(x) x^{-\alpha} \tag{6.17}$$

where $\alpha > 1$ and $L(x)$ is a slowing varying function so

$$\lim_{x \to \infty} L(tx)/L(x) = 1 \tag{6.18}$$

for t being a constant. The property of $L(x)$ follows from the requirement that $p(x)$ should be asymptotically scale invariant; thus, the form of $L(x)$ only controls the shape and finite extent of the lower tail. Scale invariance and polynomial relationship between two variables can be defined as:

$$f(x) = ax^k + o(x^k) \tag{6.19}$$

where a and k are constants, and $o(x^k)$ is an asymptotically small function of x^k. k is typically called the scaling exponent, where the word *scaling* denotes the fact that a power-law function satisfies $f(cx) \propto f(x)$ where c is a constant. Thus, a rescaling of the function's argument changes the constant of proportionality but preserves the shape of the function itself. 4

Principal component analysis

It is an orthogonal linear transformation of data in a new system such that each new variable is as less correlated as possible. Each of the new axes to describe the data, called principal components, are expressed in decreasing order of variance explanation. 51

Protein-protein interaction network (PIN)

Ensemble of proteins that relate physically and/or functionally. 2

Python

Python is a high-level programming language. It accounts for its readability and easy to learn. Programming style is not forced supporting both object-oriented and structured programming. Many scientific libraries are available and it is used by several important users as NASA or Google. 144

Runge-Kutta method

Runge-Kutta methods is a robust numerical integration methods to solve ODEs given an initial value. Different modifications have been developed since its first implementation in the 1900s. One of

the most broadly used is the fourth order Runge-Kutta consisting on the following. Taken the initial value problem,

$$y' = f(t, y), \quad y(t_0) = y_0 \tag{6.20}$$

the new value is calculated as

$$y_{n+1} = y_n + \frac{1}{6}h(k_1 + 2k_2 + 2k_3 + k_4), \tag{6.21}$$

being $h = \Delta t$ and,

$$k_1 = f(t_n, y_n) \tag{6.22}$$

$$k2 = f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1) \tag{6.23}$$

$$k3 = f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_2) \tag{6.24}$$

$$k4 = f(t_n + h, y_n + hk_3) \tag{6.25}$$

In this case, contrary to the Euler method, the slope is averaged at four points, two middle points (for which double weigh is counted; two different slopes are used to locate them) and at the beginning and at the end of the interval. 33

Shannon entropy

The Shannon entropy is a measure of the uncertainty of a random variable. Mathematically defined as,

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \tag{6.26}$$

where n is the number of elements of the random variable X and b the base of the logarithm. In the case of $b = 2$, the value of $H(X)$ is measured in bits. 17

Shannon entropy

A measure of uncertainty of a random discrete variable X. It is defined as,

$$H(X) = -\sum_{i}^{n} p(x_i) \log_b p(x_i), \tag{6.27}$$

where n is the number of possible states of X, p is the probability mass function and b is the base of the logarithm. Entropy is measured in bits if $b = 2$ or hartleys if $b = 10$. 94

Small-world network

A small-world network is a network that exhibits high clustering coefficient and low shortest path average, compared to random networks. Specifically, the average shortest path length scales at most logarithmically with the number of nodes for fixed mean degree [592]. Formally described by Watts and Strogatz in 1998 [592] it has been claimed that small-world properties can be found at many real networks, from the nematode *Caenorhabditis elegans* neural network to film actor collaboration networks. 5, 19

Sobol's method

Sobol's method is a computational method to calculate variance-based global sensitivity indexes [524]. It is based on the decomposition of the function $\boldsymbol{y} = f(\boldsymbol{x})$ into terms of higher dimensionality,

$$f(x_1, x_2, ..., x_k) = f_0 + \sum_i f_i + \sum_i \sum_{j>i} f_{ij} + ... + f_{12...k}. \quad (6.28)$$

Accordingly the variance D associated to $\boldsymbol{y}$ can be decomposed,

$$D = \sum_i D_i + \sum_i \sum_{j>i} D_{ij} + ... + D_{12...k}, \quad (6.29)$$

or

$$D = \sum_{s=1}^{k} \sum_{i_1<...<i_s} D_{i_1...i_s}. \quad (6.30)$$

Sobol's global sensitivity indexes are defined as,

$$S_{i_1...i_s} = \frac{D_{i_1...i_s}}{D} \quad (6.31)$$

which quantifies the amount of variance for the combination of parameters $i_1...i_s$ with respect to the total variance. 37, 148

SourceForge

Largest open source software development site hosting more than 230,000 software projects and over 2 millions of registered users. 144

Stiffness

Applied to differential equations, stiffness refers to numerical instability. A practical common source of stiffness in ODEs systems is a

very wide range in the parameter values: then some terms account for large variations in the solution and large errors are conducted during the integration. 33

Stochastic process
A stochastic process is a collection of random variables. The stochastic process is specified by the properties of the joint distribution for those random variables. Mathematically, we can define it as a collection $\{X_t|t \in T\}$, being T a parameter space and X a random variable. Examples of stochastic processes are random walks, Brownian motion, a Poisson process or a Markov chain. 30

Systems Biology Markup Language (SBML) [268]
Computer-readable format based on XML for representing models of biological processes, being the *de facto* standard format in computational biology [144] with over 180 software tools (July, 2010) supporting it. 53–56, 59–61

Tabu search
Tabu search is a local search optimisation method which is characterised by the *tabu list*, which is a list of once explored parameter values that are bypassed. Further references are [211, 212, 114, 213]. 11

Tandem affinity purification (TAP) experiment
Biochemical technique consisting on the addition of a TAP tag to generally the C-terminus of a protein. The TAP tag consists on a calmodulin binding peptide, a tobacco etch virus protease cleavage site and a Protein A. The process of purification consists first on a column of IgG, to which Protein A binds. Then the tobacco etch virus protease is used to cleavage part of the TAP tag and a second column of purification is used next, this time of calmodulin, for which the binding is reversible. The final purification elution holds the protein under study and its binding partners. 87

Trace of a matrix
Defined on a square matrix, it is the sum of the elements of the main diagonal. 50

Transcription factor (TF)
Molecule with capabilities to direct regulation of gene expression.

It binds to the DNA, recognising specific sequences and promoting or inhibiting the transcription of a DNA sequence into mRNA. 2

Transcriptional regulatory network (TRN)
Ensemble of genes and transcription factors with functional (regulatory) relationships. Because relations are effected by transcription factors over genes or over other transcription factors, TRNs are commonly represented by oriented graphs. 2

Vertex
As defined in [407], a vertex is the fundamental unit of a network, i.e., a person in a social network, a computer in computer networks or a protein in a protein-protein interaction network. 2

Vertex accessibility
Mean shortest distance from a given vertex to the rest of all or a part of all other vertices of the graph. 19

Yeast two-hybrid (Y2H) experiment
High-throughput experiment to elucidate physical interaction among proteins. The basic idea of the technique consists on the expression of a reporter gene by a genetic construct. The construct consists on a protein A (commonly called *bait*) fused to a DNA binding domain and a protein B (also called *prey*) fused to an activating domain. Only when proteins A and B interact, the reporter gene is expressed. 2, 87

Notes