Date of acceptance Grade

Instructor

Concepts of network biology

Ville Vahervuori

Helsinki March 26, 2007

Seminar report

UNIVERSITY OF HELSINKI
Department of Computer Science

Ville Vahervuori

Työn nimi — Arbetets titel — Title

Concepts of network biology

Oppiaine — Läroämne — Subject

Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Seminar report	March 26, 2007	12 pages

Tiivistelmä — Referat — Abstract

The cell is the common unit of structure shared among all living organisms where almost all vital processes take place. More then a century of research has been extremely successful in first breaking up the cellular components and then identifying the majority of all molecules inside cells. However, every identified molecule is just a single link in a typically very complicated chain of different chemical reactions. Therefore, these reactions have to be contemplated under a wider perspective to fully understand all the processes within a living cell. Network biology is a relatively new bioinformatical approach to model these processes and building a conceptual framework based on well-established knowledge on networks.

ACM Computing Classification System (CCS):
J.3 [Life and medical sciences]

Avainsanat — Nyckelord — Keywords

bioinformatics, network biology, cellular networks

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction **1**

2 Concepts of networks **2**

2.1 Basic measures and properties 2

2.2 Network topology 4

3 Biological networks **4**

3.1 Common features 5

3.2 Evolution of scale-free networks 5

3.3 From network structure to function 6

3.4 Identifying functional modules 7

3.5 Robustness 8

4 Properties of the links **9**

5 Conclusion **10**

References **11**

1 Introduction

Understanding the processes within a living cell and the functions of its components is such an essential topic of biology that a huge effort has been made to solve the mysteries of this very generic structure of all living organisms. For more than a century the main approach to this has been *reductionism*, where the idea is that all complex systems can be gradually reduced to a set of simpler structures. This research has been indeed very successful in identifying cellular molecules and their functions in biochemical reactions within and between cells. In the field of protein research this can be shown as nowadays complete sets of the transcripts of organisms are available, i.e. the analysis of the whole yeast transcriptome [Vel97].

However, this research has also shed light on the enormous complexity of the biochemical machinery inside cells. It has become apparent that the full understanding of the cellular processes cannot be obtained by simply identifying every molecule present in a cell. In most cases, numerous proteins work together performing a single task. To understand this task the complex interactions between many proteins have to be considered and therefore modelled in a way that is both simple and informative but still biologically correct. Linear reaction chains such as the breakdown of sugars or even more complex reactions like the well-known citric acid cycle have traditionally been depicted by graphs with arrows. Such a representation is nothing else than a network with, in these case, linear or cyclic topology. Approaching biological problems by applying knowledge from abstract network and graph theory is a current topic of bioinformatics called *network biology*.

This paper is mainly based on the work of Barabási et al. [BaO04, Jeo00] and therefore focusing on the concepts of network biology and its contribution to understanding the functional organization of cells. In addition, examples from other areas will be mentioned to show the wide applicability of general network theory. The next section will give a brief summary of general terminology and measures of networks that will be used throughout the paper. Section 3 introduces the concepts of biological networks and their specific common features. Further characteristics of links and their consequences to modelling biological systems are then covered in section 4. Finally, the paper closes with comments on network biology in general and its current status.

2 Concepts of networks

The idea of networks is so universally applicable to complex systems that it has been adopted in many fundamentally different domains such as technology, psychology and sociology. It has been even used to describe structures like the political organization of a big and socially extremely heterogeneous country like India [CoM58] or more technically, analyzing its railway system [Par03]. The Internet is probably the best example of a complex system that is strongly governed by general laws of network theory although evolving quite independently. This section covers the basic measures of graph theory, which will be later used to characterize biological networks.

2.1 Basic measures and properties

Every network consists of **nodes** and the **links** between them. The links represent some kind of relationship between two nodes. If there is no logical difference whether node A is linked to B or vice versa, the links are represented as plain lines and the network is said to be **undirected**. Otherwise, the links are arrows and the network is then **directed**. The **degree** k of a node is the amount of links associated with the node. In directed networks one has to distinguish between outgoing and incoming links, thus resulting in two degrees k_{in} and k_{out}. For example in Fig. 1 node A has k_{in}=4 and k_{out}=1. The **degree distribution** *P(k)* is obtained by counting the number of nodes with k links and dividing that by the total amount of nodes in the network. This represents the probability that a node in the given network has exactly k links. In Fig. 1 from the total of eight nodes C, G, H and E have two links, thus $P(2)= 4/8 = 0{,}5$. If we depict the distribution of all the degrees in a graph, we observe a global maximum for P(2). Clearly peaked degree distributions are one characteristic of arbitrary generated so-called *random networks* like the one depicted in Fig 1. As we will see in the next section, the degree distribution plays a major role in classifying different types of networks.

If the degree distribution approximates a power law $P(k) \sim k^{-\gamma}$ the network is said to be **scale-free**. Scale-free networks exhibit many distinct features in contrast to random networks. They have a non-uniform topology where most nodes have low degrees but few nodes, so-called *hubs*, are highly connected. The constant γ is called the **degree exponent** and determines many important properties of the underlying systems especially the importance of these hub nodes.

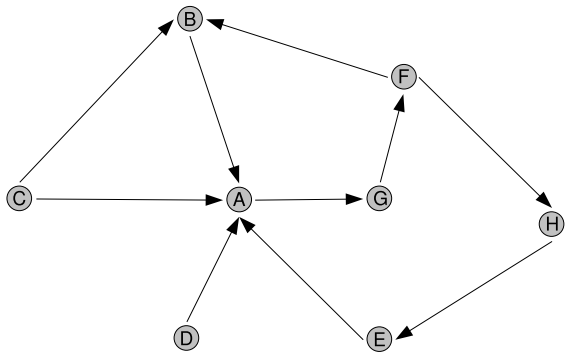

Figure 1: Directed random network

Path lengths are important network measures to describe the navigability of either specific nodes or the whole network. The **shortest path** between two nodes is simply the path with the fewest links in between. The **mean path length** is the average of the shortest paths of all pairs of nodes. Calculating these measures a bit more compicated in directed networks where typically many paths between nodes are not existing at all and the existing ones get longer (see Fig. 1: no links are leading to neither C nor D, and the shortest path from A to B goes through G and F).

One last important measure concerning also biological networks is the **clustering coefficient** C_A and its derivatives which describe whether node A is part of clusters, i.e. triangles. In Fig. 1, only nodes B and C form a triangle with A and in fact, A, B and C are the only nodes in this network being part of a triangle. C_A is the ratio of triangles going through node A and the total amount of theoretically possible triangles given that node A has k neighbours. Formally $C_A = 2n_A/\mathrm{k(k\text{-}1)}$, where n_A is the amount of links connecting node A's k neighbours with each other. For node A in Fig. 1, $k = 5$, $n_A = 1$ and thus $C_A = 2/20$. The **average clustering coefficient** $<C>$ is simply the average of the clustering coefficients and therefore characterizes the overall tendency of nodes being part of clusters. Since measures that depend on the size of the network (i.e. number of nodes/links) are not useful for characterizing different networks, a more general way to describe the clustering of nodes is the function *C(k)* which is the average clustering coefficient for all nodes with k links.

2.2 Network topology

Only by looking at the graphs of networks, we are usually already able to roughly characterize their structure. Different topologies are interesting since they typically exhibit specific features. The main three types are *random, scale-free* and *hierarchical networks*. Mathematical properties of random networks have been studied since almost half a century [ErR60] but much later the terms scale-free [BaA99] and hierarchical [Rav02] have been introduced to graph theory.

A random network topology can be generated with first placing all nodes N and then linking each pair of nodes with some probability p. This results in a seemingly random arrangement of nodes and links. Scale-free networks also look random but they contain few highly linked hubs in contrast to the majority of nodes with very few links. The result is the power-law degree distribution mentioned earlier. Hierarchical networks also may also have hubs but more important the overall structure is repeating itself in its substructures, see Fig. 2. Typical features of hierarchical networks are clustering coefficients that approximates to $C(k) \sim k^{-1}$ and a power law degree distribution like scale-free networks (more specifically, with a degree exponent $\gamma = 1 + \ln 4/\ln 3$).

3 Biological networks

In the postgenomic age of genetics where high-throughput methods generate vast amounts of information it is essential to process this data efficiently otherwise this bottleneck is keeping many important insights hidden. To get a better understanding of a cell's functional organization large numbers of interacting biochemical reactions have to be analysed. If we can model the reactions inside a cell as a complex network, we might be able to exploit the results of the already well-studied area of network theory. From here, highly efficient or even optimal solutions have been found for many very common network-related problems such as search or clustering algorithms. This section covers the properties that characterize biological and especially cellular networks.

3.1 Common features

One important feature is that most networks inside cells are scale-free. Most cellular substrates are only involved in few reactions whereas a few others, the hubs, are performing a huge number of tasks. Among proteins, which are the omnipresent substrates in metabolic pathways, there are many known *hub proteins* e.g. pyruvate or coenzyme A. Although most biochemical reactions are theoretically reversible, which would mean an undirected network, the cellular environment typically forces the reactions into one specific way making metabolic pathways generally directed networks.

Research has shown that in complex networks the shortest paths between any two nodes is amazingly short. This so-called *small-world effect* applies to social networks, neural networks, the WWW and many more. Although being a property of random networks, scale-free network exhibit this property as well. In fact, they are even *ultra small* compared to random networks [CoH03]. This was first observed in metabolic networks where paths rarely exceed three or four reactions. These short path lengths in metabolic systems allow the cell to response very quickly to environment changes.

Disassortativity is another feature of cellular networks [MaS02]. This means that hubs usually do not link directly to other hubs. A opposite scenario is often observed in social networks, where well-connected persons typically tend to know each other personally. However, no reasons for disassortativity in cellular networks have been reported yet.

3.2 Evolution of scale-free networks

If scale-free networks are so dominant in cellular networks, what are the reasons? It is widely accepted that *preferential attachment* is one major cause for the development of hubs. This means that any new node in the network tends to link to well-connected nodes making that even more connected. This is very evident in the WWW where pages prefer linking to well-known pages rather than unknown ones. Of course, the requirement for this theory is that the network is a result of constant growth. This is obviously true for most networks, especially cellular networks that have undergone millions of years of continuous evolution.

Unlike web pages, new genes are typically not simply created or introduced by some external means. For the growth of cellular networks *gene duplication* provides a reasonable explanation [RzG01]. Gene duplication is a rare but natural occurrence

where a gene is duplicated to another locus on the DNA or RNA. Being an exact copy, it will result in an identical protein, if transcribed. This new node will interact with the same partners as the protein translated from the original gene. Preferential attachment can be seen when gene duplication is observed from a node's point of view. If a node has many links and a random duplication occurs, then it is more likely that it comes from one of the own neighbours thus giving it another link and eventually it might turn into a hub. In other words, a node with only one link will gain an additional link only in the relatively rare case his sole neighbour's gene gets duplicated.

3.3 From network structure to function

Formation of groups, or *modularity* is a common feature of all sorts of networks. Circles of friends in social networks, web pages with similar topics and in cellular networks we typically find a group or *module* of functionally linked molecules that carry out one rather distinct function. Most molecules are only active when being part of a specific complex of different molecules. These complexes can be seen as functional modules and some of them are at the core of many very basic biological functions. Identifying these reappearing modules enables us to assign them an already known biological functionality.

As modules are groups of interconnected nodes a high modularity within a network can be identified by viewing at the clustering coefficient C. Theoretically, random and scale-free networks should have a similar average clustering coefficient. However, we observe significantly higher $<C>$-values for most real biological networks, ranging from metabolic to protein-protein interaction networks. Therefore, it seems that high modularity is in fact a general property of *biological* scale-free networks. In the next subsection, hierarchical clustering is introduced to explain high C-values in scale-free networks.

So far, the discussion about topology focused on the overall structure of the network. However, a bottom-up analysis often shows distinct small-scale patterns such as triangles, squares etc. inside the network. For example in Fig. 1 nodes A, B, C form a triangle whereas A, G, F, B a square. These specific arrangements are referred to as *subgraphs* and in a sufficiently complex random network all of these subgraphs are expected to be present. Interestingly, in certain networks some of these subgraphs are clearly over represented, whereas others do not exist at all [Itz03]. Indeed, some subgraphs, known as *motifs*, are responsible for specific cellular functions such

as directed triangles, so-called feed-forward loops, in transcription-regulatory and neural networks. In contrast, a feed-forward loop of four directed nodes, creating a square, is a common motif in electronic circuits but not in biological networks [Mil02]. In fact, motifs have been found inside all real complex systems so far [Mil02] and here a typical observation is, that same types of motifs tend to cluster together thus forming *motif clusters*. However, motif clustering naturally results in interaction between motifs. Understanding these motif-motif interactions is an active area of research and not yet sufficiently understood.

3.4 Identifying functional modules

The main task of identifying motifs in networks is to find subgraphs that are significantly more frequent than one would expect from a randomized network. This can be achieved by combinatorial means considering every subgraph of all n nodes. This is feasible for small networks, however the amount of possible subgraphs grows exponentially with the amount of nodes and therefore it is obviously not practical for complex systems. Another problem of module identification is the fact that the idea of relatively isolated modules is contradictory with our assumption that most biological networks are scale-free. As mentioned earlier, highly connected hubs keep the network tied closely together, thus making isolated clusters unlikely. Hierarchical networks as shown in Fig. 2 exhibit a topology that has both isolated clusters and well-linked hubs. Although not as symmetrical as depicted in Fig. 2, hierarchical organization to some extend is indeed ubiquitous in real complex systems [Rav02, RaB03].

Identifying modules in large networks seems to be a daunting task because a breakdown of the whole network into the smallest, but still biologically relevant, clusters is needed. Fortunately, here the long-lived research of networks has yielded good results. Many clustering algorithms provide automated solutions to find modules within modular networks. However, the term module is not clear-cut and different algorithms find different boundaries for the modules. If we look at the right graph in Fig. 2: is the green or the blue triangle the desired module? Not being a weakness of the algorithms, this is a logical consequence of a hierarchical structure. Most clustering algorithms provide internal parameters that adjust the preferred module sizes. The results of different runs should be compared against each other to find suitable values. Pure mathematical clustering approaches will always contain this ambiguity. The addition of existing knowledge about functional modules can address this issue

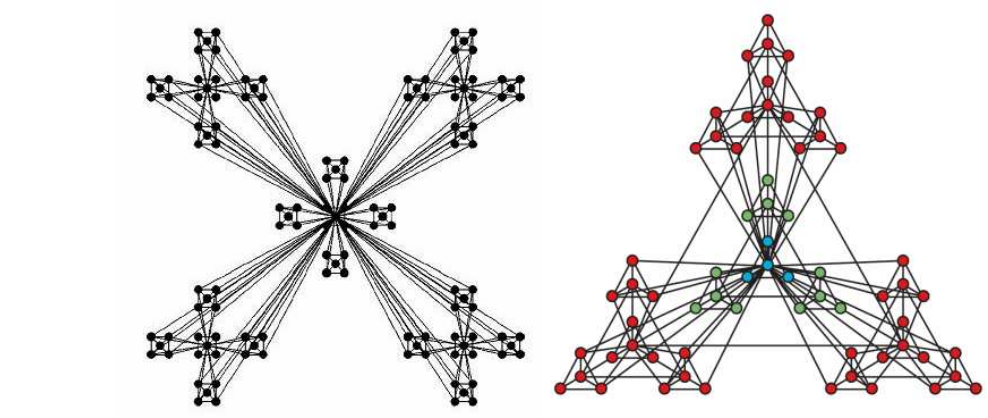

Figure 2: Hierarchical networks (sources: left [Bil07]; right [BaO04]).

and knowledge-based algorithms have already been proposed [Ihm02]. Then again, the use of hierarchical clustering algorithms in wrong contexts should be avoided since they may find hierarchical clusters even in clearly non-hierarchical networks.

3.5 Robustness

Apart from the analysis of substructures, the surprising overall robustness of complex systems has drawn attention. It is common knowledge that nodes and/or their links in all types of networks tend to become spontaneously unavailable. In computer sciences, a trivial example is a crashed server, in biology a possible reason could be a temporarily or permanently inactivated gene that results in a missing node in a metabolic network. The Internet as a fine example of a robust complex network is amazingly invulnerable to failures of nodes or changes of the operating environment. Thus, robustness defines the sensitivity of a network to modifications.

Topological robustness determines how deletions of nodes alter the overall topology of a network. Hereby the underlying topologies play a major role. Disabling a certain amount of randomly chosen nodes, a so-called critical fraction, will render a random network into a set of non-communicating islands. On the other hand, large scale-free networks do not have these values where disintegration of the network will be expected. The reason for this is that the randomized deletion will mainly address the numerous less-connected non-hub nodes. [AJB00] suggest that up to 80% of the nodes in a scale-free network can be cut off with the remaining 20% still

forming a cluster where all pairs of nodes have a path between them. Of course, this induces an *attack vulnerability* to the hubs, because disabling only a few key hubs will completely disintegrate the network. Results of deletion analysis in Baker's yeast have shown that the amount of interactions (links) a protein has, clearly correlates with the lethality of the cell when this protein is absent [Win99]. The importance of hubs can also be seen in their significant conservation during evolution. The higher the degree of a yeast cell's protein in the cellular network, the higher the probability that a closely related orthologs of this protein will be found in higher organisms.

Functional robustness of cellular networks describes the cell's ability to maintain its normal functions despite external perturbations. From a cell's point of view different nodes are not equal, because of their different cellular functions. Therefore, simply the degree of the node cannot solely determine its importance in the whole network as it is assumed from a purely topological perspective. Evolutionary aspects seem to be an important factor in functional robustness since many highly conserved metabolic modules are relatively vulnerable against modifications. On the other hand, evolutionary less conserved metabolic pathways are more robust and show ability to adapt to perturbations thus allowing the module to evolve.

4 Properties of the links

Characterization of the links in cellular networks is probably one of the most problematic topics in network biology. Biochemical reactions depicted as arrows or even simpler, plain lines are certainly crude simplifications of real cellular processes. Although easy to understand this Boolean-type of approach where a chemical reaction, exists or not, is biologically incorrect, as most reactions occur to some extend at any time and independent of state of the cell. Therefore, to obtain more precise modelling, a measure to quantify cellular processes has to be included [ScP98]. This can be taken into account with a weighted network, where the weight of a link can be for example the amount of a chemical product of a metabolic reaction, the so-called *flux*, e.g. measured in molecules per hour. This quantification of cellular processes is addressed with *metabolic flux analysis (MFA)*. However, even results of MFA are not the absolute truth because to limit logical and computational complexity they contain many assumptions (e.g. steady-state) that depending on the context may or may not be biologically correct.

Even if a Boolean approach is used, simple arrows might not provide sufficient infor-

mation. For example, modelling of gene expression gets significantly more complex by the fact that substrates that regulate the expression of a gene can exhibit positive or negative regulation. In other words, the expression of a gene can require both a high concentration of a specific substrate and at the same time the absence of another. On the other hand, positive regulation is also ambiguous. Considering two factors A and B, that both positively regulate expression of gene G. This will be depicted as two arrows, labelled A and B, leading to node G. However, this does not tell, whether both A and B are required or only one, either A or B. To cater for all these cases, additional information such as logical operators must be used not only in visualization but also in the inferred equations for the formal modelling of the process. Knowledge from the theory of Boolean networks can be applied to some extend but these results may also show clear boundaries to what can be solved computationally at all and what not.

5 Conclusion

Network biology is certainly a very welcome approach considering the increasing demand for automated methods to process the ever-growing amount of biological data extracted from existing well-automated experimental methods. The discovery of the scale-free property of biological networks caused a considerable scientific discussion and exploitation of these properties promised unforeseen results.

However, lately this excitement has clearly calmed down as concrete examples for precise modelling of complex biological systems are still missing. Even though we can draw large metabolic networks and characterize them by means of mathematical formulas, so far even the most complex processes for which reliable computable models, i.e. differential equations, have been developed do not contain much more than ten genes. This means that we are still far away from high throughput solutions to model and predict cellular processes especially processes on a large scale.

The vision of finding universal laws that govern complex biological processes is tempting but rather unrealistic. The attempt of inferring general and fundamental properties from a chosen set of previous observations is easily ill-fated, since many things can be concluded if subjectively interpreted. This is especially true for a non-exact science like biology. In most cases, the conclusion is simply that real biology is much more complicated than anticipated and large simplification is only rarely meaningful.

References

AJB00 Albert, R., Jeong, H. and Barabási, A. L., Error and attack tolerance of complex networks. *Nature*, 406(2000), pages 378–382.

BaA99 Barabási, A. L. and Albert, R., Emergence of scaling in random networks. *Science*, 286(1999), pages 509–512.

Bil07 Bilar, D., Science of Networks, 2006. `http://cs.wellesley.edu/~dbilar/cs298/ScienceofNetworks.mm.html`. [24.02.2007]

BaO04 Barabási, A. and Oltvai, Z., Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2004), pages 101–113.

CoH03 Cohen, R. and Havlin, S., Scale-free networks are ultra-small. *Phys. Rev. Lett.*, 90,058701(2003).

CoM58 Cohn, B. S. and Marriott, M., Networks and centres of integration in indian civilization. *Journal of Social Research*, 1(1958), pages 1–9.

ErR60 Erdõs, P. and Rényi, A., On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5(1960), pages 17–61.

Ihm02 Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N., Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31,4(2002), pages 370–377.

Itz03 Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G. and Alon, U., Subgraphs in random networks. *Phy. Rev. E. Stat. Nonlin. Soft Matter Phys.*, 68(2003), page 026127.

Jeo00 Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabási, A. L., The large-scale organization of metabolic networks. *Nature*, 407(2000), pages 651–654.

MaS02 Maslov, S. and Sneppen, K., Specificity and stability in topology of protein networks. *Science*, 296(2002), pages 910–913.

Mil02 Milo, R., Shen-Orr, S. S., Itzkovitz, S., Kashtan, N. and Alon, U., Network motifs: simple building blocks of complex networks. *Science*, 298(2002), pages 824–827.

RaB03 Ravasz, E. and Barabás, A. L., Hierarchical organization in complex networks. *Phys. Rev. E.*, 67(2003), pages 026112–026119.

RzG01 Rzhetsky, A. and Gomez, S. M., Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. *Bioinformatics*, 17(2001), pages 988–996.

Rav02 Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabási, A. L., Hierarchical organization of modularity in metabolic networks. *Science*, 297,5586(2002), pages 1551–1555.

Par03 Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P. A., Mukherjee, G. and S., M. S., Small-world properties of the indian railway network. *Physical review*, 67,3(2003), pages 1–5.

ScP98 Schilling, C. H. and Palsson, B. O., The underlying pathway structure of biochemical reaction networks. *Proc. Natl Acad. Sci. USA*, 95(1998), pages 4193–4198.

Vel97 Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B. and Kinzler, K. W., Characterization of the yeast transcriptome. *Cell*, 88,2(1997), pages 243–51.

Win99 Winzeler, E. A. e. a., Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(1999), pages 901–906.