Connected!
The topology of biochemical networks.

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium

an der Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

Von

Stefan Wuchty

im Mai 2002

Für Grete und Toni

THX!

Ein chinesischer Wunsch - ob positiv oder negativ intendiert - besagt, 'Mögest Du interessante Zeiten erleben!'. Diese Arbeit ist das Zeugnis interessanter Zeiten, die ich in den vergangen Jahren in Heidelberg erlebt habe. Diese Zeiten waren gespickt mit Höhen, die ich erklommen habe, aber auch genauso vielen Tiefen, in die ich, wie jede(r) DissertantIn, gestürzt bin. Glücklicherweise gibt es Menschen, die helfen, wieder aus den Tiefen zu finden. Zu jenen, die mich immer geund bestärkt haben, in dunklen Stunden nicht aufzugeben, zählen meine Eltern, Margarethe und Anton, meine Schwester Margit und meine Arbeitsgruppenleiterin Ursula Kummer.

Ohne die Inspirationen, die mir Ursula geschenkt hat, wäre die Qualität der Arbeit nicht möglich gewesen. Mit (doktor)väterlicher Nachsicht und Güte ist mir immer Peter Schuster begegnet, der die offizielle und meine manchmal nicht ganz leichte Betreuung übernommen hat.

Die Geschäftsleitung, Klaus Tschira und Andreas Reuter, ermöglichten das Umfeld am European Media Lab.

Albert-László Barabási und Erik Sonnhammer luden mich zu Vorträgen an der Notre Dame University in den USA und am Karolinska Insitutet in Stockholm ein.

Ursula Rost und Ralph Gauges halfen mir oft genug aus rechnerspezifischen Problemen. Besonders Ralph erwies sich als sehr begeisterungsfähig für Wienerisches, ebenso wie Bill Andersen und Jennifer Williams. Der Ex-Kärntner Stephan Jansen sorgte dafür, daß der für Wiener so wichtige Schmäh nie ausging. Bärbel Mack, Kornelia Gorisch und Claudia Spahn erleichterten das Überqueren so mancher organisatorischer Hürde. Zusammen mit Ursula Kummer, Ursula Rost, Ralph Gauges, Stephan Jansen, Alexandra Martin, Renate Kania, Dennis Pfisterer, Christian Elting, Katja Wegner, Norbert Rabes, Achim Beck, Beate Keller, Carel van Gend, Jürgen Zobeley, Isabel Rojas, Luca Bernardi, Esther Ratsch, Christoph Müller, Rainer Malaka, Kelly Elkins, Timm Essigke, Razif Gabdoulline, Kerstin Schneider, Rebecca Wade, Reinhold Weinmann und vielen anderen sorgten sie für eine entspannte und freundschaftliche Athmosphäre am EML.

Abstract

Real life systems were recently found to demonstrate scale-free and small-world behavior instead of random graph characteristics. In this study, the topology of some biological networks is discussed.
Protein domain networks are studied which treat domains as nodes and their co-occurrence in sequences as edges. It is found that these networks exhibit small-world and scale-free topologies with a high degree of local clustering accompanied by a few far distance connections. Moreover, these observations apply not only to complete databases but also to the domain distributions in proteomes of different organisms. The extent of connectivity among domains reflects the evolutionary complexity of the organisms considered.
Henceforth, data of currently available protein-protein interaction sets and protein domain sets of *Saccharomyces cerevisiae* are used to set up protein and domain interaction and domain sequence networks. In a comparison, they all turn out to be sparse and locally well clustered indicating scale-free and partially small-world topology. Frequently, triangles of connected nodes are found in these topologies which represent short cuts in the networks. Their amount is measured by a newly introduced transitivity coefficient. Fairly small sets of highly connected proteins and domains shape the topologies of the underlying networks emphasizing a kind of backbone the nets are based on. The biological nature of these particular nodes is further investigated. Since highly connected

proteins and domains accumulated a significant higher number of links by their important involvement in certain cellular aspects, their mutational effect on the cell is considered by a perturbation analysis. A comparison of the domain and domain interaction networks of *Saccharmoyces cerevisiae* considers the factors which force domains to accumulate links to other domains in protein sequences of higher eukaryotes.

The abundance of sequenced genomes enabled the investigations of genomic homogeneity and heterogeneity emphasizing segments of varying lengths. Network based approaches to provide a common perspective on nucleotide segments are presented. A segmentation of sequences to equal lengths results in adjacent pieces. So, these segments are treated as nodes and their mutual adjacency as edges. For comparison purposes, analyses are carried out on exon and intron sequences as well as on randomly generated sets of equal size. Regardless of the origin of sequences, a transition from Gaussian distributions of connectivity to real power-laws which indicate scale-free topology towards increasing segment sizes is observed. Relative entropy is applied as a divergence measure to networks which were set up by natural sequences and their corresponding randomly sampled pendants revealing that network topologies differ strongly over a very narrow range of segment sizes. The tendency that segments preferably occur either in exons or introns with increasing size indicates the opportunity to use them as probes for the detection and investigation of exons and introns. Consequences from a biological and evolutionary perspective and conclusions with regard to already published works are discussed.

Finally, conformational spaces of a tRNA set up by suboptimal structures and a move set which defines transitions between them are discussed from the perspective of small-world networks. Henceforth, the enhancing influence of modifications on typical small-world network properties and the shape of energy landscapes thus obtained is considered.

Zusammenfassung

In den letzten Jahren wurde mehrmals gezeigt, daß reale Netzwerke eher Charakteristika von sogenannten Scale-Free- und Small-World-Netzwerken besitzen als von Zufallsgraphen. In der vorliegenden Arbeit wird die Topologie von einigen biochemischen Netzwerken untersucht.
Zuerst werden Netzwerke betrachtet, die Proteindomänen als Knoten und deren gemeinsames Auftreten in Proteinsequenzen als Kanten behandeln. Diese Netze zeigen starke Scale-Free- und Small-World-Eigenschaften, die sich durch ein hohes lokales Clustering auszeichnen. Diese Cluster sind durch wenige weitreichende Kanten untereinander verbunden. Diese Beobachtungen gelten nicht nur für komplette Datenbanken von Proteindomänen, sondern auch für die Proteome von einzelnen Organismen. Der Grad, wie stark Domänen untereinander verbunden sind, spiegelt die Komplexität des zugrundeliegenden Organismus wieder.
Als Weiterführung dieser Ideen werden Netzwerke von *Saccharomyces cerevisiae* verglichen, die auf Proteininteraktionen, Proteindomäneninteraktionen und den soeben erwähnten Proteindomänen basieren. Alle diese Netzwerke zeigen Charakteristika von Scale-Free-Netzen. Teilweise können auch Eigenschaften von Small-World-Netzwerken gefunden werden. Diese Topologien basieren auf Dreiecken von Knoten, die als Abkürzungen ('short cuts') betrachtet werden können. Diese Eigenschaft wird von einem neuen Transitivitätskoeffizienten gemessen. Allen diesen Netzwerken liegt zugrunde, daß deren Topologien von einigen weni-

gen stark verknüpften Proteinen und Domänen dominiert werden, die eine Art Rückgrat bilden. Die biologische Natur dieser Knoten wird weiter untersucht. Da stark verknüpfte Knoten eine deutlich höhere Anzahl von Kanten durch deren bestimmenden Einfluß auf zelluläre Aspekte besitzen, wird deren Effekt auf die Störungsanfälligkeit der Netzwerke untersucht. Mit einem Vergleich der Netzwerke von Domäneninteraktionen und Domänen wird untersucht, welche Einflüsse Domänen veranlassen, Verbindungen zu anderen Domänen zu akkumulieren.
Sequenzierte Genome ermöglichen die Untersuchungen der genomischen Homo- und Heterogenität mit Segmenten unterschiedlicher Länge. Um eine einheitliche Perspektive auf Sequenzsegmente zu ermöglichen, werden in dieser Arbeit netzwerkbasierende Ansätze behandelt. Betrachtet man Sequenzsegmente gleicher Länge als Knoten und deren unmittelbare Lage zueinander als Kanten, ergeben sich interessante Topologien. Um eine Vergleichsmöglichkeit zu haben, werden diese Netzwerke aus Segmenten sowohl von Exons, Introns als auch von völlig zufällig generierten Sequenzen aufgebaut. Es stellt sich heraus, daß unabhängig vom Ursprung der Sequenzen mit wachsender Länge der Segmente ein Übergang von einer Gaußverteilung der Verbindungen zu einem Power-Law, das auf ein Scale-Free-Netzwerk deutet, in Verteilungen der Kantenanzahl pro Knoten festgestellt werden kann. Als Maß für die Divergenz der Topologien von Netzwerken, die aus Segmenten von natürlichen und zufälligen Sequenzen bestehen, wird die relative Entropie verwendet. Es zeigt sich, daß starke Unterschiede in den Netzwerktopologien nur über einen engen Bereich von Segmentlängen existieren. Schließlich wird die Tendenz der Segmente, eher zu einem bestimmten Satz von Exons oder Introns zu gehören, behandelt. Diese Tendenz eröffnet die Möglichkeit, Segmente als Sonden zur Detektion und Untersuchung von Exons und Introns einzusetzen. Diese Resultate werden von der Warte bereits veröffentlichter Arbeiten diskutiert.
Suboptimale Strukturen einer tRNA und ein Move-Set, das die Übergänge zwischen den Strukturen definiert, ermöglichen die Darstellung des zugrundeliegenden Strukturraums. Graphen, die die Konformationen der tRNA behandeln, erweisen sich als Small-World-Netze. Der verstärkende Einfluß von Modifikationen auf die typischen Eigenschaften der Small-World-Netzwerke und auf die zugrundeliegenden Energielandschaften wird diskutiert.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The study of various networks is currently pervading all of science, ranging from physics to biology. The most important issues are the structures of the networks. How can one characterize the connections of the Internet or the metabolic network of *E.coli*? Are there any fundamental principles which made the networks emerge? How does the procedure of setting up networks reflect features of evolution especially in biological systems? Thus, we would like to understand how a tremendous network of interacting entities like proteins, metabolites or domains behave collectively given their coupling architecture. Currently, we witness the begin of unraveling the structure and henceforth the functions and properties of complex biological networks.

Often the connection topology of networks was assumed to be either completely regular or random (Erdös and Rényi 1960). However, these network types proved to be ill suited to describe the topology of many real and especially biological systems. Thus, it is necessary to conceive new ways to model topologies in order to explain features of networks better.

Watts and Strogatz revealed a new class of network topologies that lies some-

where between these two extremes. Originally, these small-world networks were generated by randomly rewiring nodes in a regular network. Small-world networks combine the local clustering of connections characteristic of regular networks with occasional long-range connections between clusters, as can be expected to occur in random networks. By defining measures that distinguish these three types of networks, the authors showed that several biological, technological and social networks are of the small-world type (Watts and Strogatz 1998). A small-world graph is formally defined as a sparse graph which is much more highly clustered than an equally sparse random graph (Bollobás 1998). Small-world graphs were first illustrated with friendship networks (Milgram 1967) in sociology, often referred to as 'six degrees of separation' (Guare 1990). The architecture of the power grid of the western United States, the structure of some sociological networks dealing with mathematical collaborations on publications, and the casting of actors in movies were found to be small-world graphs (Watts and Strogatz 1998).

Barabási et al. introduced a theoretical model that generates graphs demonstrating a connectivity distribution which decays as a power-law. This feature was found to be a direct consequence of two generic mechanisms:

I. Networks are allowed to expand continuously by the addition of new vertices.

II. These newly added nodes attach preferentially to sites that are already well connected (Barabási and Albert 1999).

Since this feature is independent of the actual size of the network, they called this class of inhomogeneous networks scale-free networks. The topology of the World Wide Web was investigated by considering HTML documents as vertices which are connected by links pointing from one page to another (Albert et al. 1999; Barabási and Albert 1999; Barabási et al. 2000). The latter net, as well as the Internet which emerges from connecting different servers, demonstrate scale-free properties. Both nets display a high degree of robustness against errors (Albert

et al. 2000; Albert and Barabási 2002). However, these networks are highly vulnerable to perturbations of the highly connected nodes.

Recently, scale-free and small-world behavior have also been found in biological networks. Watts and Strogatz reported the architecture of the *C.elegans* nervous system to show significant small-world behavior (Watts and Strogatz 1998). Fell and Wagner assembled a list of stoichiometric equations that represent the central routes of the energy metabolism and small-molecule building block synthesis in *E.coli* (Fell and Wagner 2000). A substrate graph was constructed defined by a vertex set consisting of all metabolites that occur in the network. Two metabolites were considered to be linked if they occur in the same reaction. Most recently, Jeong et al. comparatively analyzed metabolic networks of organisms representing all three domains of life (Jeong et al. 2000). The metabolic network is represented by nodes, the substrates, connected by directed edges symbolizing the actual reaction. The topology of these networks are best described by a scale-free model. Furthermore, the diameters of the nets remain the same for all these networks regardless of the number of substrates found in the given species. Interestingly, the ranking of the most connected substrates is largely identical for all organisms. These highly linked nodes dominate the topology which are suggested by the scale-free model as the immediate result of repeated preferential attachment. Considering the latter procedure as a rough abstraction of evolution, the set of most connected substrates was treated as an evolutionary core. Like the technical networks, the *E.coli* network theoretically has high tolerance to random errors but severe sensitivity towards the removal of the highly connected nodes.

In this study, several biological networks will be investigated towards the emergence of these topologies.

Protein domain networks generated with data from the ProDom, Pfam, Prosite and InterPro domain databases is studied. It is found that these networks exhibit small-world and scale-free topologies with a high degree of local clustering accompanied by a few far distance connections. Moreover, these observations

apply not only to the complete databases but also to the domain distributions in proteomes of different organisms. The extent of connectivity among domains reflects the evolutionary complexity of the organisms considered.

Subsequently, data of currently available protein-protein interaction sets and protein domain sets of *Saccharomyces cerevisiae* are used to set up protein and domain interaction and domain sequence networks. All of them are far from being random or regular networks. In fact, they turn out to be sparse and locally well clustered indicating scale-free and partially small-world topology. Frequently, triangles of connected nodes are found in these topologies which represent short cuts in the networks. Their amount is measured by a newly introduced transitivity coefficient. Fairly small sets of highly connected proteins and domains shape the topologies of the underlying networks emphasizing a kind of backbone the nets are based on. The biological nature of these particular nodes is further investigated. Since highly connected proteins and domains accumulated a significant higher number of links by their important involvement in certain cellular aspects their mutational effect on the cell is considered by a perturbation analysis. A comparison of the domain and domain interaction networks of *Saccharmoyces cerevisiae* considers the factors which force domains to accumulate links to other domains in protein sequences of higher eukaryotes.

Domain networks were found to display these topologies which immediately suggested their evolutionary meaning. So, it might be interesting to also investigate networks which are set up by genomic sequence segments. A segmentation of sequences to equal lengths results in adjacent pieces. These segments are treated as nodes and their mutual adjacency as edges. For comparison purposes, analyses are carried out on exon and intron sequences as well as on randomly generated sets of equal size. Regardless of the origin of sequences, a transition from a Gaussian-shaped degree distribution to a power-law towards increasing segment size is observed. Applying relative entropy as a measure of divergence, topologies of natural and corresponding random segments prove to vary significantly over a narrow range of segment sizes. The tendency of segments to occur preferably either in exons or introns indicates the opportunity to distinguish exons from

introns by relatively short segments. Similarly to the domain networks, the degree of connectivity displays the evolutionary complexity of the organism. Consequences from a biological and evolutionary perspective and conclusions with regard to already published works are discussed.

Finally, a closer look to the conformational space of a typical tRNA is taken which is set up by sets of suboptimal structures from the perspective of small-world networks. The enhancing influence of modifications on typical small-world network properties and shapes of energy landscapes is discussed.

This study is organized as follows: Chapter 2 reviews relevant network topologies providing information which are necessary to understand the conclusions based on the results of the upcoming parts. The other chapters are organized in the order they were addressed above. Since the latter chapters are intended to be self-containing, they can be read independently from each other. Finally, chapter 7 considers some future ideas.

CHAPTER 2

Connected! - Network topologies

Growing amounts of empirical, theoretical and especially biological data about the topologies of large complex networks indicate the emergence of several network types. Basically, these types are classified by the connectivity distribution of nodes, $P(k) \sim k^{-\gamma} e^{k/\xi}$, where $e^{k/\xi}$ denotes a cut-off at some characteristic scale ξ. Three classes can be defined (Amaral et al. 2000):

I. If ξ is very small, $P(k) \sim e^{k/\xi}$. Thus, the distribution is single scaled. Typically, this distribution would correspond to a Gaussian or exponential distribution. Prominent protagonists of this type are the random graph model (Erdös and Rényi 1960) and the small-world model (Watts and Strogatz 1998). Both lead to fairly homogenous networks with nodes comprising approximately the same number of links $k \sim \langle k \rangle$ (Barabási et al. 1999). Small-world graphs adopt sparse topologies but remain more highly clustered than equally sparse random graphs (Watts and Strogatz 1998).

II. Provided that ξ grows, a power-law with a sharp cut-off is obtained;

III. If ξ is large, a proper power-law, $P(k) \sim k^{-\gamma}$, occurs which is typical for scale-free networks (Barabási and Albert 1999). Compared to exponential

networks, the probability that a node is highly connected ($k \gg \langle k \rangle$) is statistically significant in scale-free networks (Barabási and Albert 1999).

In the following, the main types of network topologies will be reviewed. Special emphasis will be put on their set-up, features and relevance for biological systems.

2.1 Erdös-Rényi model

The most frequently investigated random network model was first introduced by Erdös and Rényi (Erdös and Rényi 1960; Bollobás 1998). This model starts with N vertices and no edge. With a distinct probability p, each pair of vertices is connected which leads to the emergence of a random network. Interestingly enough, many features of this network type appear quite suddenly at a threshold value $p(N)$. A significant property of this topology is the emergence of trees and cycles. A tree of order k is a connected graph with k vertices and $k-1$ edges. A cycle of order k is a cyclic sequence of k edges, so that every consecutive edge has a common vertex. It has been demonstrated that almost all vertices belong to isolated trees if $p \sim c/N$ with $c < 1$. Abruptly, cycles of all orders appear at $p \sim 1/N$. The Erdös-Rényi model proves to be important in percolation analyses. In this context, $p_c \sim 1/N$ is the percolation threshold of a system. Significantly, the system breaks into many small clusters for $p < p_c$. At p_c, large cluster form which contains all vertices in the asymptotic limit.

In the Erdös-Rényi model, the probability that a vertex has k edges follows the Poisson distribution

$$P(k) = e^{-\lambda} \frac{\lambda^k}{k\,!}, \tag{2.1}$$

where

$$\lambda = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{2.2}$$

It is easy to show that increasing numbers of p lead to broader distributions. The expectation value of the Poisson distribution proves to be $(N-1)p$.

2.2 Small-world model

2.2.1 The model

The description of a transition from a locally ordered system to a random network revealed an interesting new topology that is often referred to as the small-world model which was first observed by Watts and Strogatz (Watts and Strogatz 1998).

The algorithm is based on a two level process:

I. *Start with order*: It starts with a number of N vertices which are initially linked to its adjacent and next-adjacent neighbors (Generally, the model could also include neighbors up to an order of n).

II. *Randomize* : Subsequently, each vertex is rewired with probability p which means that one end of the vertex is shifted to an other randomly chosen vertex. This rewiring process has to meet some criteria. First, no two vertices are allowed to share more than one edge and, second, no vertex is allowed to have an edge with itself.

The graphs considered are sparse but not so sparse that they would become disconnected. Specifically, graphs have to fulfill the qualification

$$N \gg k \gg \ln(N) \gg 1 \tag{2.3}$$

where $k \gg \ln(N)$ guarantees that a random graph does not get disconnected (Bollobás 1998).

The structural properties of the networks are quantified by two parameters. The mean path length $\langle L \rangle$ is defined as the number of edges in the shortest path between two vertices averaged over all pairs of vertices. The mean clustering coefficient is defined as follows: Provided that node i has k_i neighbors, then at most $k_i(k_i - 1)/2$ edges can exist between them which obviously proves if all k_i neighbors are also connected to each other. Thus, C determines the fraction of

allowable links which indeed exist by

$$C_i = \frac{2E_i}{k_i(k_i - 1)}, \tag{2.4}$$

where E_i is the number of edges which actually exists between these k_i nodes. $\langle C \rangle$ is defined as the average of C over all i. Obviously, L determines global properties while C measures the cliquishness of a typical neighborhood.

Obviously, a probability $p = 0$ would leave the graph regular while $p = 1$ would convert the networks topology to a random one. In this regime, it was found that

$$L \sim n/2k \gg 1 \ , \ C \sim 3/4 \text{ if } p \to 0, \tag{2.5}$$

while

$$L \approx L_{random} \sim \ln(N)/\ln(k) \ , \ C \approx C_{random} \sim k/n \ll 1 \text{ if } p \to 1. \tag{2.6}$$

So, the regular graphs emphasize a high clustered topology leading to linearly growing mean path lengths between pairs of nodes. In contrast, random networks provide a poor clustered topology and show a logarithmic dependency of the mean path length.

It was found that in the interval $0 < p < 0.01$ the model exhibits so-called small-world properties. The network thus obtained is sparse and preserves approximately the same mean path lengths through the network,

$$\langle L \rangle \geq \langle L \rangle_{random}, \tag{2.7}$$

but prove to be much more clustered than a random graph of equal size,

$$\langle C \rangle \gg \langle C \rangle_{random}. \tag{2.8}$$

2.2.2 Features

The connectivity distribution of the small-world model depends strongly on p. $p = 0$ leads to distributions $P(k) = \delta(k - z)$, where z is the coordination number of the lattice. Finite p generates distributions which still peak around z but

prove to be broader. With $p \to 1$, the connectivity distribution approaches to that obtained for the random graph model with $p = z/N$. It has been shown that the connectivity distribution of the small-world model for $p > 0$ and large N corresponds to

$$P(k) = \sum_{n=0}^{f(k,K)} C_{K/2}^{n}(1-p)^{n} p^{K/2-n} \frac{(pK/2)^{k-K/2-n}}{(k-K/2-n)!} e^{-pK/2} \tag{2.9}$$

for $k \geq K/2$, where $f(k, K) = \min(k - K/2, K/2)$ (Barrat and Weigt 2000; Albert and Barabási 2002). In this equation, K denotes the average degree of nodes and C^n the clustering coefficient of respective n nodes. As already mentioned, the shape of the degree distribution is similar to that of a random graph. It has a pronounced peak at $\langle k \rangle = K$ and decays exponentially for large k. The topology is thus mainly homogenous considering nodes which have approximately the same number of nodes.

The average path length $\langle L \rangle$ reveals some interesting properties. The small-world model immediately suggests a drastical change in $\langle L \rangle$ which depends on increasing values of the fraction p of rewired edges. For small p, $\langle L \rangle$ scales linearly while for large p a logarithmic dependency is observed. The reason for the rapid drop is the appearance of shortcuts between nodes. Thus, this shortcuts connect even remote parts of the network resulting in significantly decreased path lengths. Apparently, a so-called crossover length which depends on p might exists. Numerical simulations and analytical considerations suggest that the mean path length obeys the general form

$$\langle L \rangle (N, p) \sim \frac{N}{K} f(pKN^{d}), \tag{2.10}$$

where $f(u)$ is a universal scaling function that is constant if $u \ll 1$ and is $\ln(u)/u$ if $u \gg 1$. d denotes the dimension of the original lattice (Barrat and Weigt 2000; Barthélémy and Amaral 1999; Newman and Watts 1999; de Menezes et al. 2000; Watts 1999).

The dependence of $C(p)$ can be estimated using a slightly different definition

of the clustering coefficient

$$C' = \frac{3 \times N_{triangles}}{N_{connected\,triples}}. \tag{2.11}$$

Here triangles are trios of nodes in which each node is connected to both of the others . Connected triples refer to trios in which at least one is connected to both others . Thus, the factor 3 corresponds to the fact that each triangle contributes to 3 connected triples. To calculate $C'(p)$, consider the regular lattice which exhibits $C(0)$. For $p > 0$, two neighbors of a node i that were initially connected at $p = 0$ remain neighbors of i and connected by an edge with probability $(1-p)^3$ since there are 3 edges which have to remain intact (Barrat and Weigt 2000; Newman et al. 2001) . Thus,

$$C'(p) \simeq C(0)(1-p)^3. \tag{2.12}$$

It has been verified that the deviation of $C(p)$ from this expression is small and goes to 0 as $N \to \infty$ (Barrat and Weigt 2000).

2.2.3 Examples

Small-world graphs were first illustrated with friendship networks (Milgram 1967) in sociology, often referred to as 'six degrees of separation' (Guare 1990). The architecture of the power grid of the western United States, the structure of some sociological networks dealing with mathematical collaborations on publications, and the casting of actors in movies were found to be small-world graphs (Watts and Strogatz 1998; Newman 2001; Newman et al. 2001; Amaral et al. 2000).

Recently, small-world behavior has also been found in biological networks. Watts and Strogatz reported the architecture of the *C.elegans* nervous system to show significant small-world behavior (Watts and Strogatz 1998). Fell and Wagner assembled a list of stoichiometric equations that represent the central routes of the energy metabolism and small-molecule building block synthesis in *E.coli* (Fell and Wagner 2000; Wagner and Fell 2001). A substrate graph was constructed defined by a vertex set consisting of all metabolites that occur in the network.

Two metabolites were considered to be linked if they occur in the same reaction. They found the substrate graph to be sparse with glutamate, coenzyme A, 2-oxoglutarate, pyruvate and glutamine having the highest degree of connectivity. This sample of metabolites might be viewed as a core of *E.coli* metabolism which was found without any subjective criteria.

Most recently the conformation space of a lattice protein was found to exhibit small-world topology (Scala et al. 2001) . Conformations refer to nodes which are linked if both structures are able to switch to each other by an elementary move.

2.3 Scale-free model

2.3.1 The initial model

Commonly, both network models described above emphasize connectivity distributions, $P(k)$, which provide an exponential cut-off and a characteristic size, $\langle k \rangle$, depending on the respective p. However, many real-life systems have the common property that $P(k)$ is free of scale over broad orders of magnitude. Strikingly, the models previously discussed emphasize a constant number of nodes, N. It is clear, however, that many systems are not limited to this assumption. In fact, they are dynamic and show a constant increase of the total number of nodes, N, throughout the life time of the system. Consequently, a common feature of real life systems is the continuous expansion of the networks by the addition of newly introduced nodes which attach to already existing ones.

A second incompatibility of the network models described previously is the uniform probability, p, that two vertices are connected. In contrast, real systems tend to link nodes preferentially. Thus, the probability, p, that vertices are linked is not uniform but exhibit a higher probability of a newly introduced node to be attached to an already existing, well connected one.

Thus, the emergence of scale-free network topology which leads to a scale invariant connectivity distribution, $P(k)$, is essentially based on two major points:

I. *Growth:* Starting with a small number of nodes, m_0, at every time step a new node will be added which is allowed to set m ($\leq m_0$) edges.

II. *Preferential attachment:* The choice of the node to which the newly introduced one will be connected depends on the connectivity, k_i, of this particular node. Thus,

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}. \tag{2.13}$$

After t time steps the model leads to a random network with $N = m_0 + mt$ nodes and mt edges. Henceforth, the networks evolves into a scale-invariant state. The connectivity distribution proves to be a power law, $P(k) \sim k^{-\gamma}$ with $\gamma \approx 3$. Obviously, the scaling exponent is either independent of m, t and subsequently of the network size, $N = m_0 + t$. Hence, the network organizes itself into a scale-free stationary state despite its continuous growth. A schematic survey of the ideas described so far is given in Figure 2.1.

The combination of growth and preferential attachment leads to interesting dynamics of the individual vertex connectivities. The vertices which provide the most links are those that entered the network at an early stage since vertices grow proportionally to their connectivity relative to the rest of nodes. Since some of the oldest nodes had a long time to aquire links, they are responsible for the high-k part of $P(k)$.

In order to get an analytical result for the connectivity of a particular vertex, Barabási et al. suggested a mean-field approach (Barabási et al. 1999).
If k is continuous the probability to attach newly introduced nodes preferentially, $\Pi = k_i / \sum_j k_j$, can be treated as a continuous rate of change of k_i. Thus,

$$\frac{\partial k_i}{\partial t} = A\Pi(k_i) = A\frac{ki}{\sum_{j=1}^{m_0+t-1} k_j}. \tag{2.14}$$

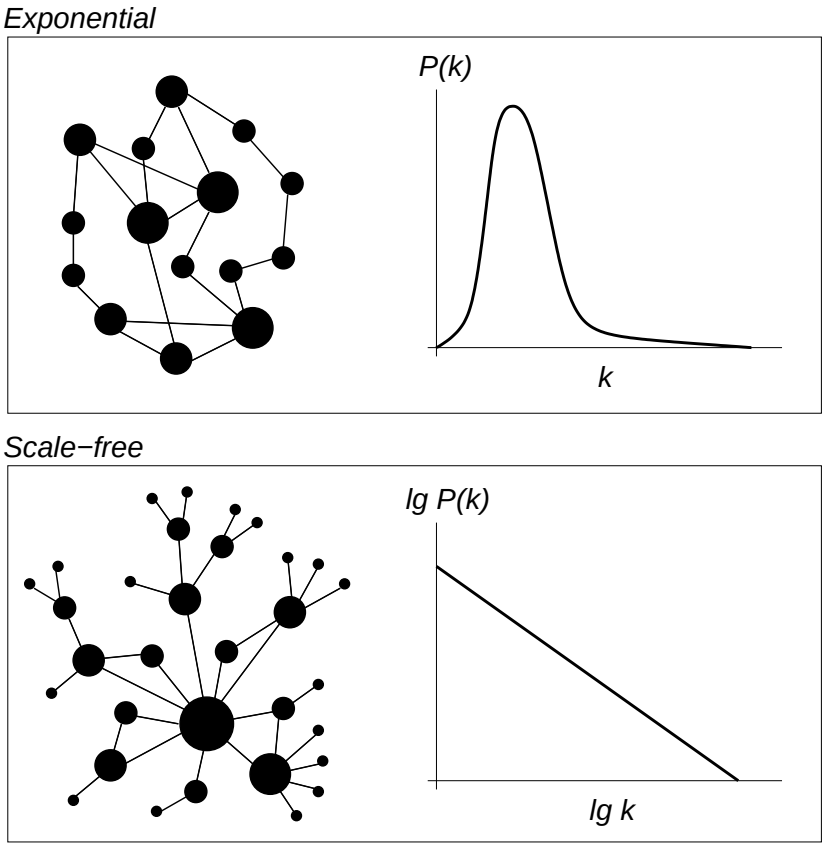


Figure 2.1: Models of exponential and scale-free networks. Diameters of circles indicate the number of connections respective nodes have. $P(k)$ is the frequency that nodes have k connections. Top: Exponential networks consist of nodes which show similar numbers of links to other nodes. Thus, the frequency distribution peaks at an average and decays exponentially. Bottom: In fact, biological networks adopt scale-f ree topology. A fairly small amount of highly connected nodes which show much higher numbers of connection than the aver age shapes a straight line in the log-log plot of the connectivity distribution.

Recalling that $\sum_j k_j = 2mt$ and the change in connectivities at a time step is $\Delta k = m$, it follows that $A = m$. Thus,

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}. \tag{2.15}$$

The solution of this equation provided that vertex i was added to the system at time t_i with connectivity $k_i(t_i) = m$, is

$$k_i(t) = m\sqrt{\frac{t}{t_i}}. \tag{2.16}$$

Obviously, older (i.e. smaller t_i) nodes increase their connectivity at the expense of the younger (i.e. larger t_i) nodes leading to highly connected nodes. Such

a 'rich-gets-richer' phenomenon can easily be detected in real life systems. The latter expression can easily be used to calculate the power-law exponent γ analytically. Thus, the probability that a vertex has connectivity $k_i(t)$ which is smaller than k, $P(k_i(t) < k)$ can be written as

$$P(k_i(t) < k) = P\left(t_i < \frac{m^2 t}{k^2}\right). \tag{2.17}$$

Provided that vertices are add uniformly at each time increment to the system, the probability density of t_i is

$$P_i(t_i) = \frac{1}{m_0 + t}. \tag{2.18}$$

Substituting into equation 2.14 it follows that

$$P\left(t_i > \frac{m^2 t}{k^2}\right) = 1 - P\left(t_i \leq \frac{m^2 t}{k^2}\right) = 1 - \frac{m^2 t}{k^2(t + m_0)}. \tag{2.19}$$

The probability density for $P(k)$ is obtained using

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{2m^2 t}{m_0 + t}\frac{1}{k^3}, \tag{2.20}$$

suggesting $\gamma = 3$ which is in good agreement with the experimental results obtained. Obviously, γ proves to be independent of m.
The latter expression additionally predicts the coefficient of the power-law distribution, $P(k) \sim Ak^{-\gamma}$, to be proportional to m^2, ie. $A \sim m^2$.

2.3.2 Features

Real networks provide the coexistence of clustering and short path lengths. Thus, it is interesting to note if the scale-free model contributes some small-world properties.

It can be shown empirically that the mean path length$\langle L \rangle$ of the scale-free model depends on the network size in a logarithmic manner. Thus,

$$\langle L \rangle = A \log(N - B) + C \tag{2.21}$$

as a generalized logarithmic form applies to this distribution. Although there is no analytical expression which provides a good estimate of the path length in

scale-free models there were some first steps which emphasized the n-loop structure of the underlying networks (Gleiss et al. 2001).

The same conclusion holds for the clustering coefficient $\langle C \rangle$. Similarly to the decay in random graphs, $\langle C_r \rangle = \langle k \rangle / N$, the clustering coefficient decreases albeit slowlier in the same manner. However, there is also no analytical expression of $\langle C \rangle$ available (Albert and Barabási 2002).

2.3.3 Examples

There is a bunch of technological, social and biological systems scale-free topology appears in. The topology of the World-Wide Web was investigated by considering HTML documents as vertices which are connected by links pointing from one page to another (Albert et al. 1999; Barabási and Albert 1999; Barabási et al. 2000; Huberman and Adamic 1999; Kleinberg and Lawrence 2001; Lawrence and Giles 1998). The latter net, as well as the Internet which emerges from connecting different servers, demonstrate scale-free properties. Both nets display a high degree of robustness against errors (Albert et al. 1999; Albert et al. 2000; Amaral et al. 2000). However, these networks are highly vulnerable to perturbations of the highly connected nodes.

Social networks such as citation and collaboration networks as well as the web of human sexual contacts turn out to be scale-free (Liljeros et al. 2001; Vazquez 2001; Barabási et al. 2002).

Most recently, Jeong et al. comparatively analyzed metabolic networks of organisms representing all three domains of life (Jeong et al. 2000). The metabolic network is represented by nodes, the substrates, connected by directed edges symbolizing the actual reaction. The topology of these networks are best described by a scale-free model. Furthermore, the diameters of the nets remain the same for all these networks regardless of the number of substrates found in the given species. Interestingly, the ranking of the most connected substrates is largely identical for all organisms, thus indicating hubs which dominate the topology

of the nets. Like the technical networks, the *E.coli* network theoretically has high tolerance to random errors but severe sensitivity towards the removal of the highly connected nodes.

Also protein-protein interaction networks display this topology. Jeong et al. showed that the degree distribution of the physical protein interaction map of Yeast follows a truncated power-law with an exponential cut-off $P(k) \sim (k+k_0)^{-\gamma} e^{-(k+k_0)/k_c}$ with $k_0 = 1, k_c = 20$ and $\gamma = 2.4$.

2.3.4 Limitations of the scale-free model

Despite the good results obtained so far, some brief thoughts should be taken about limiting cases of the scale-free model. Clearly, continuous growth and preferential attachment are necessary for the emergence of a power-law scaling. However, what will happen if both ingredients work separately from each other? Consider a network which keeps the growing character of scale-free network but attaches newly introduced nodes uniformly. Thus, the model is defined as:

I. *Growth:* Starting with a small number, m_0, of nodes, at every time step a new node will be added which is allowed to set m ($\leq m_0$) edges.

II. *Uniform attachment:* Every new node connects with equal probability to the vertices which are already present in the system, i.e. $\Pi(k_i) = 1/(m_0 + t - 1)$

Analogously to the procedure already introduced, the connectivity distribution takes the form

$$P(k) = \frac{e}{m} \exp\left(-\frac{k}{m}\right), \tag{2.22}$$

with e being a constant.
Thus, the absence of preferential attachment eliminates the occurrence of the typical power-law (Barabási and Albert 1999).

The second model tests the opposite scenario:

I. *No growth:* The total number of nodes, N, in the system remains constant.

II. *Preferential attachment:* At each time step, one node is selected randomly which is connected with the probability $\Pi = k_i / \sum_j k_j$ to node i in the system.

Since N is constant and the number of edges increases in time, all vertices are connected at least after $T \simeq N^2$ time steps.
The time-evolution of the individual connectivities can also be calculated using the mean field approach stating that

$$k_i(t) \simeq \frac{2}{N} t. \tag{2.23}$$

Obviously, this result indicates that after a transient time of duration, $t \simeq N$, the connectivity increases linearly with time. Thus, the mean-field approximation predicts that after a transient period the connectivities of all vertices have the same value given by equation 2.23. Thus, the connectivity distribution $P(k)$ changes from an initial power-law to a Gaussian distribution as time increases.

Obviously, the loss of continuous growth and preferential attachment are necessary for the emergence of scale-free topology (Barabási and Albert 1999).

2.3.5 Further works on scale-free graphs

It is easy to see that the initial scale-free model has its limitations if it is compared to real life systems. It predicts a power-law degree distribution with a fixed exponent. However, most real networks provide exponents which vary between 1 and 3. Furthermore, the degree distribution might show a truncated power-law or a saturation for small k. Thus, it is necessary to mention briefly some interesting approaches resulting in different scaling exponents which might be more suitable to describe real systems.

Measures of the preferential attachment of real systems which is clearly the main ingredient for the emergence of scale-free properties revealed that rather a nonlinear preferential attachment, $\Pi(k_i) \sim k_i^{\alpha}$, applies to the Internet, citation

and collaboration networks (Albert and Barabási 2002). An analytical calculation resulted in a complete disappearance of scale-free properties for $\alpha \neq 1$ (Krapivsky et al. 2000). Contrasty, asymptotically linear preferential attachment, $\Pi(k_i) \sim a_\infty k$, as $k \to \infty$ indicates a power-law exponent $\gamma \to 2$ if $a_\infty \to \infty$ and $\gamma \to \infty$ if $a_\infty \to 0$ (Krapivsky et al. 2000). Similarly, $\Pi(k_i)$ might also be dependent of an initial attractiveness, considering $\Pi(k_i) \sim A + k_i$. It turns out that $\gamma = 2$ if $A = 0$ and $\gamma \to \infty$ if $A \to \infty$ (Dorogovtsev and Mendes 2000a; Dorogovtsev and Mendes 2000b).

An approach which suits the demands of networks better and extends the initial model by incorporating new edges between existing nodes and the rewiring of edges is discussed in (Albert and Barabási 2000). m new edges are added to the system with probability p, m edges are rewired with probability q and finally $1 - p - q$ denotes the probability to add a new node to the system. After some analytical calculations, the power-law exponent turns out to be $\gamma = 2$ if $q = (1 - p + m)/(1 + 2m)$ and $\gamma \to \infty$ if $p, q, m \to 0$. This approach was applied successfully to the degree distribution of the World Wide Web.
It is clear that real life systems also demand the disappearance of nodes. Accordingly, mechanisms were conceived to fulfill this requirement too (Dorogovtsev and Mendes 2000c).

Also network topologies were already considered which emphasize different fitness values of nodes. Thus, preferential attachment modifies to $\Pi(k) \sim \eta_i k_i$ resulting in a connectivity distribution $P(k_i) \sim k^{-1-m}/\ln k$, where m is a constant(Bianconi and Barabási 2001).

Obviously, there is a growing amount of different approaches to describe networks which provide a power-tail in their connectivity distributions. A good in depth survey about different types of emerging scale-free networks can be found in (Albert and Barabási 2002).

CHAPTER 3

Domain networks

3.1 Introduction

It was already noted that scale-free and small-world behavior also occurs in biological networks. In metabolism, the network topologies introduced illustrated their significance for evolutionary phenomenons. In this chapter, a biochemical network is formed by sets of domains which are linearly arranged in protein sequences. This might generate graphs comprising interesting features. Since the topology of graphs thus generated is still unknown it is worth considering this way of treating domain architectures.

3.2 Domain organisation

Protein crystallography reveals that the fundamental unit of protein structure is the domain. Independent of neighbouring sequences, this region of a polypeptide chain folds into a distinct structure and mediates biological functionality (Janin and Chothia 1985). Most proteins contain only one single domain (Doolittle

1995). Some sequences appear as multi-domain proteins adopting different linear arrangements of their domain sets. On average, such domain architectures comprise two to three domains; however, some human proteins contain up to 130 domains (Li et al. 2001).

Similar to the discussion about the role of certain metabolites in the emergence of metabolism, there has been a debate about the actual number of existing domains and their origin. One view treats all past and present proteins as the result of shuffling a large set of primordial polypeptides (Dorit and Gilbert 1991). These are assumed to result from splicing events involving exons separated by introns (Gilbert and Glynias 1993). The other view deals with the existence of a few small polypeptides in early stages of life; these are the predecessors of most contemporary proteins (Doolittle 1995). Gene duplication and subsequent modification were employed to form the latter molecules from this small set of polypeptides. Independent of the timing for the introduction of introns, recombination in introns provides a mechanism for the exchange of exons between genes. This mechanism for the acquisition of new functions by eukaryotic genes is commonly known as 'exon shuffling'. It was assumed that primitive proteins were encoded by exons that were spliced together (Seidel et al. 1992). However, such shuffling events take on biological significance only if the exons involved carry a functional or structural domain. Although many examples of exon shuffling have been found, no significant correspondence between exons and units of protein structure has been detected (Stoltzfus et al. 1994).

It is common to find that newly sequenced proteins are homologous to some other known proteins over parts of their lengths. Thus, most proteins may have descended from relatively few ancestral types. The sequences of large proteins often show signs of having evolved by the joining of preexisting domains in new combinations. Such a mechanism is commonly known as 'domain shuffling' and appears as two types: domain duplication and domain insertion (Doolittle 1995). Domain duplication refers to the internal duplication of at least one domain in a gene. Domain insertion denotes the process by which structural or functional domains are exchanged between proteins or inserted into a protein. Shuffling of

domains has more biological significance than exon shuffling because domains are real structural and functional units in proteins while exons are often not.

Functional links between proteins have also been detected by analysing the fusion patterns of protein domains. Two separate proteins A and B in one organism may be expressed as a fusion protein in other species. A protein sequence containing both A and B is termed a Rosetta Stone sequence. However, this framework only applies in a minority of cases (Marcotte et al. 1999).

3.3 Protein databases

Currently, there is a large variety of databases each collecting protein domain information in completely different ways.

The Prosite database (available at http://expasy.proteome.org.au/prosite/) consists of biologically significant motifs and profiles determined and formulated with appropriate computational tools. Uncharacterised proteins are assigned to certain protein families by the aid of weight matrices and profiles (Hofmann et al. 1999). The majority of Prosite documentation refers to motifs thus providing combined motif and domain information. Release 16.0 of Prosite contains 1374 different patterns, rules and profiles.

Another database is Pfam (http://www.sanger.ac.uk/Software/Pfam/index.shtml) which is a large collection of multiple sequence alignments of protein families and profile hidden Markov models (Bateman et al. 2000). Moreover, Pfam contains curated documentation for all 2478 families in version 5.5 covering nearly 65% of Swiss-Prot release 38 and SP-TrEmbl release 11.

Much more protein families are however found in the database ProDom (available at http://www.toulouse.inra.fr/prodom.html) (Corpet et al. 2000) which contains all protein domain families that can be generated automatically from the Swiss-Prot and TrEmbl sequence database (Bairoch and Apweiler 2000). Expert-

validated families are extended by using Pfam seed alignments to build new ProDom families with the Psi-Blast database searching algorithm (Altschul et al. 1997). Other families are generated by recursive usage of Psi-Blast. ProDom version 99.2 has 157648 domain families covering almost 95% of Swiss-Prot release 37 and TrEmbl release 10. ProDom comprises higher coverage than Pfam. However, ProDom tends to over predict the number of protein families which can be discovered as subsets of larger families.

Finally, InterPro (available at http://www.ebi.ac.uk/interpro) (Apweiler et al. 2001a) is an integrated documentation resource of protein families, domains and functional sites rationalising the complementary efforts of the Prosite, Pfam, ProDom and Prints (Attwood et al. 2000) database projects. InterPro contains manually curated documentation and diagnostic signatures from these databases and uses these to create a unique, non-redundant characterisation of protein families, domains and functional sites.

3.4 Proteome databases

The advent of fully sequenced genomes of various organisms has facilitated the investigation of proteomes. The Proteome Analysis database (available online at http://www.ebi.ac.uk/proteome) (Apweiler et al. 2001b) has been set up to provide comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms. The analysis is mainly compiled using InterPro and CluSTr (Kriventseva et al. 2001) and is performed on the non-redundant complete proteome sets of SWISS-PROT and TrEMBL entries. The latest release provides 41 non-redundant proteomes of genomes of archaea, bacteria and eukaryotes.

Most recently, SWISS-PROT and Ensembl have prepared a complete non redundant human proteome set consisting of 30585 sequences. It is the combination of the SWISS-PROT/TrEMBL non redundant human proteome set (15691 sequences) and additional non-redundant peptides predicted by Ensembl (14894

sequences). Ensembl (http://www.ensembl.org) provides complete and consistent annotation across the human genome.

In the following, domain networks generated with data from ProDom, Pfam and Prosite domain databases will be presented. Furthermore, InterPro domain networks of different species that are generated with complete proteome sets provided by the Proteome Analysis database will be considered. Subsequently, the topology of these networks will be investigated and biological and evolutionary consequences discussed.

3.5 Material and Methods

A domain graph $G_D = (V_D, E_D)$ is formally defined by a vertex set V_D consisting of all domains found within proteins. Two domains are regarded as being adjacent if they occur together in one protein at least once. An undirected edge connecting these two vertices indicates this relationship. Such connections define the edges set E_D. This graph will be investigated towards the emergence of small-world and/or scale-free properties. Thus, the degree k of a vertex is the number of other vertices to which it is linked. The mean path length L from a vertex to any other vertex of the graph is defined as the average of the path lengths to all other vertices. Another important quantity is the clustering coefficient $C(v)$ of a vertex v. It measures the fraction of the vertices connected to v which are also connected to each other. In extension, the clustering coefficient C of the graph is defined as the average of $C(v)$ over all v.

In this study protein domain information was retrieved from the ProDom, Prosite and Pfam databases. 65% of all ProDom sequences correspond to families containing 10 or more members. In order to restrict the size of the network, the sample of ProDom domains focuses on these families. Thus, 5995 ProDom domains were obtained. The Prosite database declares false negative entries which were filtered out of the sample used for the network construction. Sequence entries of each database provide Swiss-Prot annotation. Thus, every protein sequence was

itemised with each domain that it contains. This was done for each database separately. Domains which were listed due to their occurrence in one protein sequence represent vertices which are connected to each other in the domain graphs.

Complete proteome data sets of different species were retrieved from the Proteome Analysis database which uses InterPro annotation of protein domains. Such proteome data sets adopt Swiss-Prot, TrEMBL, TrEMBLnew and Ensembl annotation of proteins. Analogously, InterPro domains which appear along with other ones in a protein sequence represent vertices which are connected to each other in the domain graphs. The numbers of links to other domains in such graphs were logarithmically binned and frequencies thus obtained. Such pairs of values were subjected to a linear regression procedure.

PAJEK (the Slovene word for spider), a program for large network analysis and visualisation, was used for the calculation of the latter values (Batagelj and Mrvar 1998) (available at http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

3.6 Results

The domain graphs are sparse with small average degrees (Table 3.1) compared to the maximal possible degree $k = n - 1$ where n is the number of vertices. In this respect, the results of Figure 3.1 are interesting. The vertices which denote Prosite domains were ranked by their frequency of their connectivity. The curve is similar to a generalised Zipf's law curve in which it is observed that the frequency of occurrence of some event $f(x)$ as a function of the rank x is a power-law function $f(x) = a(b+x)^{-c}$ with the exponent c close to unity. The plot of Prosite domains in Figure 1 satisfies the latter condition with $c = 0.89$. We are thus dealing with relatively few highly connected domains and many rarely connected ones. Essentially, the frequency distributions of ProDom and Pfam domains are similar. However, they fit the generalised Zipf's law less well. Distributions following Zipf's law have also been observed in the context of literary vocabulary (Zipf 1949; Miller and Newman 1958), frequency of secondary structures of RNA

(Schuster et al. 1994), lattice proteins (Bornberg-Bauer 1997) and hits per web page in the World-Wide Web (Huberman et al. 1998). This observation is in accordance with the picture of scale-free networks which are topologically dominated by a few highly connected hubs.

	ProDom	Pfam	Prosite
n_v	5995	2478	1360
$\langle k_v \rangle$	2,33	1,12	0,77
$n_{conn.comp.}$	1394	1396	809
$n_{unconn.dom.}$	975	1316	577

Table 3.1: Some basic data of the ProDom, Prosite and Pfam graph

As illustrated in Figure 3.2, frequency distributions of vertices with degree k follow a distribution comparable to a power-law distribution. Although the shape of the distribution curves are different they share an area of linearity. Regarding these latter areas the frequency distribution of links from ProDom domains follows $P(k) \approx k^{-\gamma}$ with $\gamma = 2.5$. By contrast, the distributions of degrees of Pfam and Prosite domains follow the same law with $\gamma = 1.7$. Although the curves do not follow exactly the proposed curvature of the frequency of degrees in the original scale-free model one can observe a type of scale-free dependence even if the scale-free model is a raw approximation of the real situation. Obviously, the topology of such domain graphs is better described by a highly heterogenous scale-free or small-world model than by an exponential model.

In Table 3.2 it can be observed that the domain graphs partially satisfy the structural properties of small-world graphs. While clustering coefficients C_v of the domain graphs by far exceed the respective coefficients of corresponding random graphs, the characteristic path lengths L_v do not accomplish the demanded qualifications of a small-world graph. Emphasising the observation that the vast majority of proteins contains only one domain (Marcotte et al. 1999), the domain networks contain a huge amount of unconnected vertices (see Table 3.1).

This feature of domain distribution among protein sequences illustrates in par-

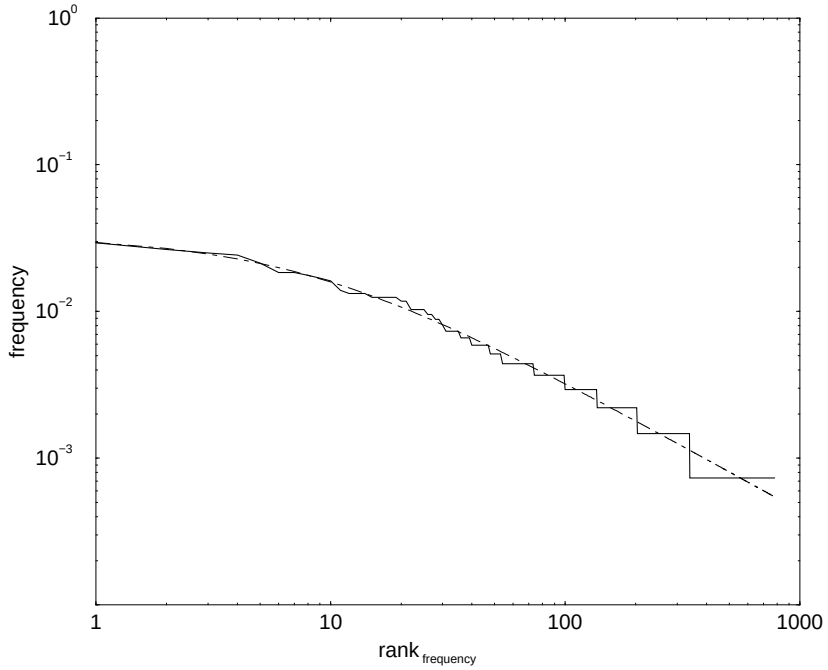


Figure 3.1: The frequency distribution of Prosite domain connectivity. The number of links to other domains are ranked by their frequencies, which follow a generalised Zipf's law: $f(x) = a(b+x)^{-c}$ with x being the rank and $f(x)$ its frequency. Parameter values of the best fit (dot-dashed curve) are $a = 0,21, b = 7,93$ and $c = 0,89$.

ticular the high number of connected components in domain graphs. Although domain graphs are thus highly scattered, every graph contains a major subnet among its connected components which gathers the majority of domains. These major components feature values of L_v and C_v values that satisfy the demand of small-world graphs by exceeding the respective values of random graphs of equal size. Thus, this study focuses on the analysis of the major components exhibiting small-world and scale-free behaviour. In order to clarify the graph topology, Figure 3.3 displays the major component of the network which was generated by

	L_{actual}	L_{random}	C_{actual}	C_{random}
ProDom	4.96	5.81	0.51	0.0008
Pfam	4.54	9.05	0.15	0.0003
Prosite	5.44	6.46	0.33	0.0044

Table 3.2: Mean path length L and mean clustering coefficient C of ProDom, Pfam and Prosite domain nets

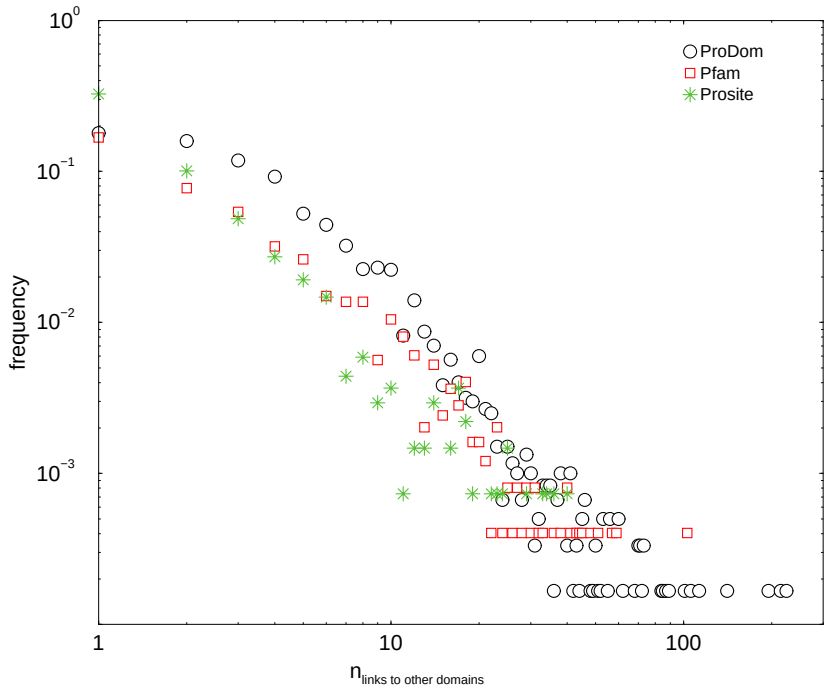


Figure 3.2: The frequency distribution of domain connections within protein sequences. Domain data were obtained from ProDom, Pfam and Prosite protein database.

proteome data of *Saccharomyces cerevisiae*.

The investigations carried out so far consider all domains without taking care of their origin. Presumably, the degree of connectivity is different if one focuses on different species. All domain connections of 6 species which developed differently in the course of evolution were extracted from the complete proteome sets provided by the Proteome Analysis database. As illustrated in Figure 3.4, the frequency distributions of links regarding *Human*, *C. elegans*, *Drosophila*, *Yeast*, *E.coli* and *Methanococcus* still follow the expected power-law. However, the slopes of the lines are slightly different. Interestingly, the slopes of *Human* and *Drosophila* nearly coincide in Figure 3.4. Moreover, the regression lines show almost the same interception in comparison to *C.elegans*. In Figure 3.4 the situation changes slightly. While the slopes in comparison to *Human* are significantly steeper, the regression lines of *Yeast*, *E.coli* and *Methanococcus* run nearly parallel. Thus, it is tempting to assume a trend which guides multicellular organisms to higher domain connectivity.

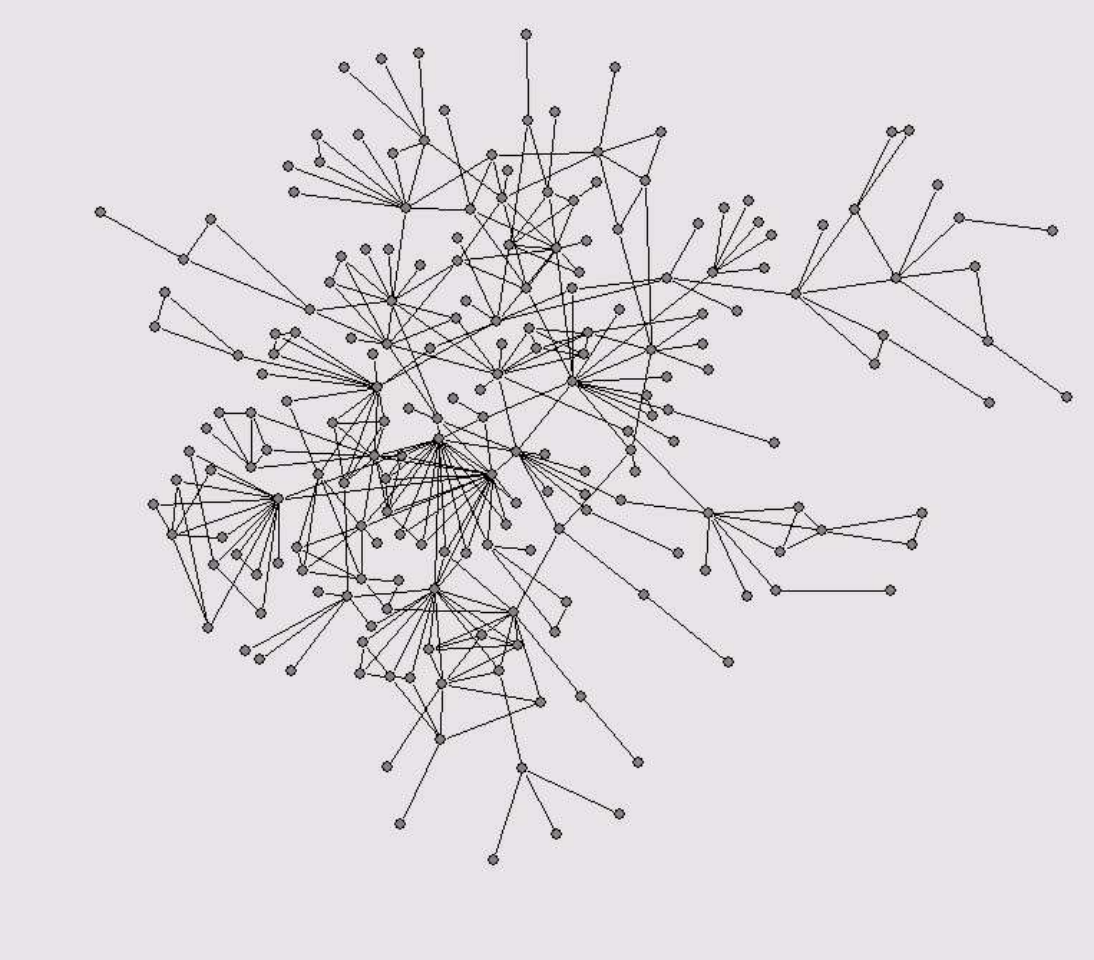


Figure 3.3: Major component of the domain network of *Saccharomyces cerevisiae* comprising 204 vertices and 347 edges

Interestingly, the majority of highly connected InterPro domains appear in signalling pathways as the list of the 10 best linked domains of different species in Table 3.6 reveals. Obviously, the evolutionary trend towards compartmentalisation of the cell and to multicellularity demands a higher degree of organisation. Therefore, more emphasis is put on the maintenance of inter- and intracellular signalling channels, cell-cell contacts and integrity. Hence, proteomes have to provide protein sets which cover such cellular demands. The growing number of highly linked domains of signalling and extracellular proteins by comparing archae, prokaryotes and eukaryotes confirms this assumption.

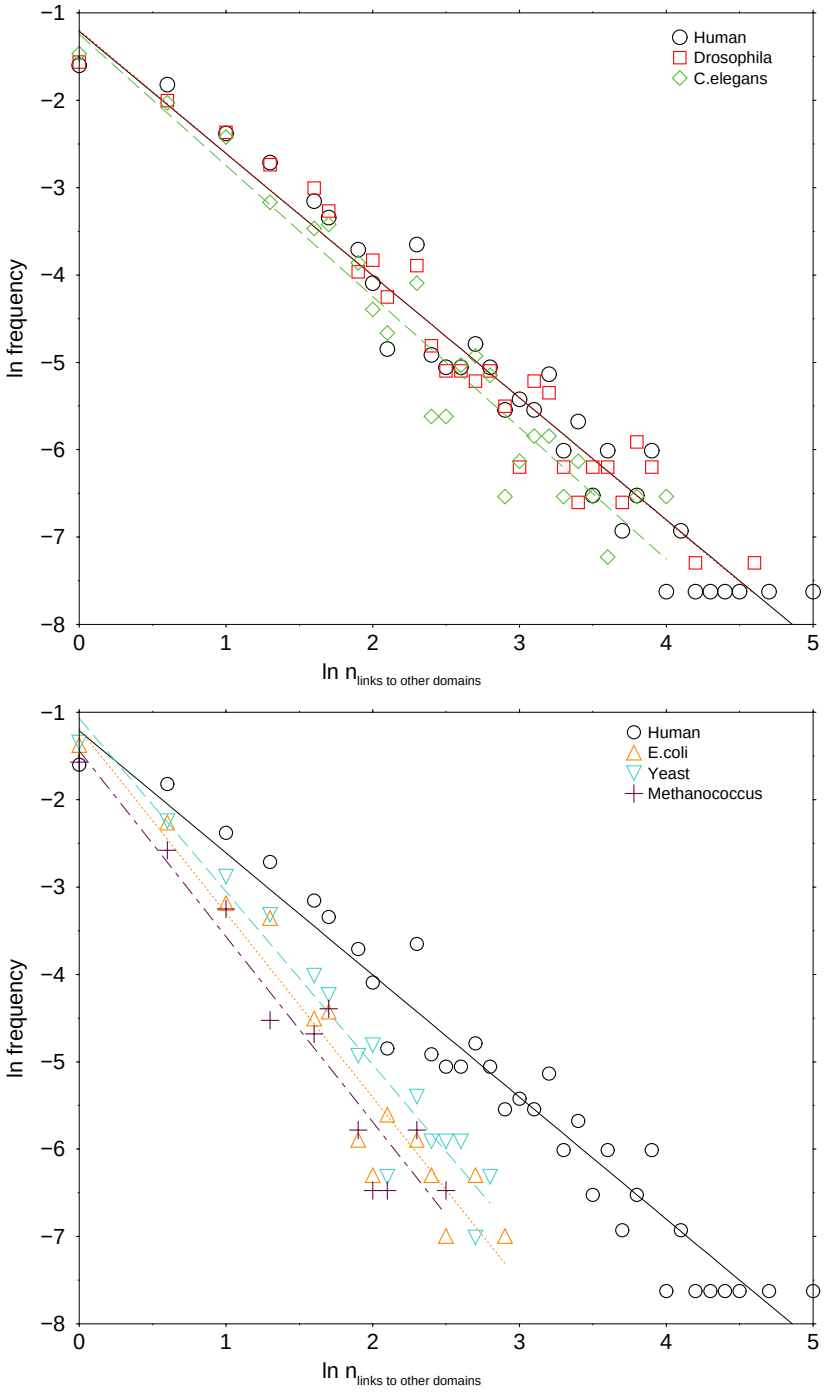


Figure 3.4: The frequency distribution of domain connections within protein sequences of *C.elegans*, *Drosophila* and *Human* (upper panel) and *Methanococcus*, *E.coli*, *Yeast* and *Human* (lower panel). The domain data were obtained from Proteome Analysis database. The numbers of links to other domains were logarithmically binned and frequencies thus obtained. These pairs of values were subject to a linear regression procedure. Regression lines of *Drosophila* and *Human* coincide.

Methanococcus		*E.coli*		*Yeast*	
domain	k_v	domain	k_v	domain	k_v
SAM	13	NAD-BINDING	20	pkinase	18
fer4	11	ESTERASE	16	P-KINASE-ST	18
FMN-ENZYMES	10	SAM	15	PH	16
NAD-BINDING	9	fer4	13	zf-C3HC4	14
AA-TRNA-LIG-I	8	AA-TRNA-LIG-II	12	AA-TRNA-LIG-II	14
intein	7	FMN	12	efhand	14
pyr-redox	7	HIS-KIN	11	C2	13
ATP-GTP-A	6	AA-TRNA-LIG-I	11	CPSase-L-chain	13
CBS	6	HIS REC	10	GATase	13
N6-MTASE	6	PAS	9	WD40	12
C.elegans		*Drosophila*		*Human*	
domain	k_v	domain	k_v	domain	k_v
pkinase	57	PRICHEXTENSN	101	ATP-GTP-A	169
EGF	57	pkinase	70	GPCRRHODOPSN	162
PH	46	zf-C2H2	53	PRICHEXTENSN	110
efhand	45	ank	52	EGF	98
ank	37	EGF	50	pkinase	89
P-KINASE-ST	35	SH3	48	ig	79
EGF-CA	34	ANTIFREEZEI	46	PH	72
zf-C3HC4	33	efhand	45	efhand	64
ig	30	PH	45	SH3	61
SH3	30	P-KINASE-ST	44	zf-C2H2	58

Table 3.3: 10 most highly connected InterPro domains of *Methanococcus E.coli*, *Yeast*, *C.elegans*, *Drosophila* and *Human*.

3.7 Discussion

What might be the functional, phylogenetic or bioinformatic implications of the power-law distribution of the connectivity of domains and the small-world behaviour of the domain networks studied?

3.7.1 Completeness and quality of data

Regardless of whether Pfam, Prosite or ProDom domain information is used, the qualitative topology of domain networks remains unchanged. Since these databases differ significantly in size and methodology, the argument is tempting that even though the current domain data are far from complete, the topology of domain networks will not change significantly with the growing amount of domain data. This assumption is supported by the characteristics of scale-free networks leading to domain graphs which are independent of the actual size of the underlying networks. Hence, the major observation that the topology of domain graphs is mainly dominated by few highly linked domains will not be changed entirely with the incorporation of new protein domain data. InterPro gathers and streamlines mostly distinct domain information from the above mentioned domain databases providing a centralised annotation resource to reduce the amount of duplication between the database resources. Hence, scale-free characteristics of InterPro domain networks which were generated with the aid of complete proteomes of different species do not change significantly in comparison to networks generated with domain information from a single database. However, it should be noted that the acquisition of protein domain information is biased to a certain extent since eukaryotic and mammalian proteins are far better studied and documented in databases on average than archeae or prokaryotic proteins.

Another import consideration regards aspects of acquisition of proteome information. Proteome data which were entirely extracted by genome translation might not explain sufficiently the set up of all cellular processes. Domain networks were generated with the aid of translated genome databases which do not cover effects that include alternative splicing and domain usage. Alternative pre-mRNA splic-

ing is an important mechanism for regulating gene expression in higher eukaryotes (Smith et al. 1989). By recent estimates, the primary transcripts of $\sim$ 30% of human genes are subject to alternative splicing. Thus, the connectivity of domains found in higher eukaryotes might be significantly higher than it is 'in silico'.

In addition, the differences in frequency distributions between higher eukaryotes, bacteria and archae in Figure 3.4 might also be related to the number of domain architectures that were found in the different organisms. Since eukaryotes and mammals developed much more distinct domain architectures (International Human Genome Sequencing Consortium 2001), the respective distributions of domain connections are statistically more reliable than those of prokaryotes and archae. Therefore, future studies should clarify, if the low number of domain architectures leads to slight artefacts in the slope of prokaryotic and archeal organisms.

3.7.2 Evolutionary aspects

Are the observed topologies the direct consequence of domain evolution? The model of Barabási and Albert generates scale-free networks by preferential attachment of newly added vertices to already well connected ones. Consequently, Fell and Wagner argued that vertices with many connections in a metabolic network were metabolites originating very early in the course of evolution (Fell and Wagner 2000) and which shape a core metabolism. Analogously, highly connected domains could also have originated very early. If one compares the lists of the most highly linked domains in Table 3.6 this assumption does not hold. The majority of more highly linked domains in *Methanococcus* and *E. coli* are mainly concerned with the maintenance of metabolism. Given that in *Methanococcus* and *E. coli* nearly none of the highly linked domains in the higher organisms can be found, and vice versa, the focus of domain connection shifts to domain hubs involved in signal transduction, transcription and cell-cell interactions. In addition, helicase C has roughly similar degrees of connections in all organisms. However, the ankyrin repeat motif (ank) is one of the few domains which can be found to be unlinked in *E. coli*, whereas it possesses a growing degree of connec-

tivity in higher eukaryotes.

Apparently, the majority of highly connected domains seems to have arisen late in eukaryotes of larger proteome size. The evolutionary trend towards multicellularity requires proteomes which feature new and additional complex cellular processes like signal transduction or cell-cell contacts. One way of accomplishing growing demands is the expansion of already existing protein sets. Indeed, many protein families are expanded in *Human* relative to *Drosophila* and *C.elegans*. These are mainly involved in inter- and intracellular signalling pathways, apoptosis (Aravind et al. 2001), development, immune and neural functions (International Human Genome Sequencing Consortium 2001; Venter, J., M. Adams, E. Myers et al. (271 co-authors) 2001). Although many protein families of these organisms exhibit great disparities in abundance, C2H2-type zinc finger motifs and eukaryotic protein kinase (pkinase) are among the top 10 most frequent domain families (Rubin, G., M. Yandell, J. Wortmann et al. (52 co-authors) 2000; Tupler et al. 2001) and of the best connected domains in Table 3.6. At least in higher eukaryotes both domains tend to increase their connections to other domains in a similar way to the already mentioned ankyrin repeat motif (ank).

Although the human phenotypic complexity exceeds the respective ones of *C. elegans* and *Drosophila* by far, proteome dimensions remain considerably low. Thus, combinatorial aspects of domain arrangements might have a major impact on the preservation of cellular processes. Among chromatin-associated proteins, transcription factors and especially apoptosis proteins, a significant portion of protein architecture is shared between *Human* and *Drosophila*. However, substantial innovation in the creation of new protein architectures was significantly detectable (International Human Genome Sequencing Consortium 2001). Apparently, expansion of particular domain families and accompanying evolution of complex domain architectures from presumably preexisting domains coincides with the increase of organism's complexity. In this regard the different slopes in Figure 3.4 indicate this evolutionary trend to higher connectivity of domains (e.g. pkinase, SH3 and EGF in Table 3.6) as well as a growing complexity in the arrangement of domains within proteins. In comparison to non-eukaryotic,

Drosophila developed more complex domain architectures. Thus, the frequency distribution of the latter organisms can be clearly separated in Figure 3.4, where lower complexity in domain architecture is indicated by steeper slopes. The first point is well reflected by the slightly different slopes of *Human*, *Drosophila* and *C.elegans* in Figure 3.4.

In conclusion, a variety of arguments point to an increase in the complexity of the proteome from the single-celled yeast to multicellular vertebrates such as *Human*. Essentially, the expansion of protein families coincides with the increase of connectivity of the respective domains. Extensive shuffling of domains to increase combinatorial diversity might provide protein sets which are sufficient to preserve cellular procedures without dramatically expanding the absolute size of the protein complement. Hence, the relatively greater proteome complexity of higher eukaryotes and especially human cannot be simply a consequence of genome size but, to a certain extent, of innovations in domain arrangements. Thus, highly linked domains represent functional centres in various different cellular aspects. They could be treated as evolutionary hubs which help to organise the domain space by occasionally linking them to numerous other functionally related domains.

3.7.3 Quality of the basic models

The view that new protein architectures can be created by shuffling, adding and deleting domains, resulting in new proteins from old parts, is well reflected by the emergence of such domain hubs. However, there exist a variety of domain arrangements which contradict the ideal image of continuous addition of new domain links to already existing hubs in the sense of scale-free networks. The S1 RNA binding domain is linked to helicase C in *E.coli*, while it is found to be connected to RNB, KH domain and RNAse PH in *Human*. Neither the procedure of generating a small-world graph in the original sense nor the scale-free model provide the deletion of vertices. However, the assumption that domains emerge and dissappear occasionally is a basic demand of protein evolution. Thus, scale-free and small-world models can obviously only be a rough approximation to the

real situation.

CHAPTER 4

Interaction and domain networks of Yeast

4.1 Introduction

Tremendous amount of biological data currently available emphasize the necessity to investigate the mutual relationships of genes, proteins and metabolites. The latter were the starting point of considering metaboloms of prokaryotes as complex networks (Fell and Wagner 2000; Wagner and Fell 2001; Jeong et al. 2000). Quite similarly, proteomes offer an opportunity to examine domain architectures of their protein sequences from this perspective (Wuchty 2001; Apic et al. 2001). Furthermore, efforts were made to enlighten interactions between families of protein domains. Structural domain data were mapped to a network linking interacting domain structures (Park et al. 2001). Finally, protein-protein interaction networks emerged by employing sets of protein interactions of *H. pylori* (Rain et al. 2001) and *S. cerevisiae* (Ito et al. 2000; Uetz et al. 2000; Ito et al. 2001; Schwikowski et al. 2000; Jeong et al. 2001).

In this chapter, the use of protein interaction data to generate an interaction network of *Saccharomyces cerevisiae* will be reported. Using the known nonredundant complete Yeast proteome, domain information is used to set up domain

sequence and domain interaction networks. Since a comparison of these three types of networks is currently undone, the topologies of these networks will be comparatively studied and biological consequences discussed.

4.2 Materials and methods

4.2.1 Definition of networks

A protein-protein interaction graph, G_{p-p}, is defined by a set of nodes which contains a set of interacting Yeast proteins. In order to complete the definition of the network protein-protein interactions are denoted by a set of undirected edges.

In a coarse grained way, a protein sequence can be computed as a linear arrangement of the domains it contains. Thus, a domain sequence graph, G_D, is formally defined by a set of nodes consisting of all domains which occur in the protein sequences of the Yeast proteome. Two domains are regarded as being undirectedly linked if they co-occur in one of these protein sequences (Wuchty 2001).

Since $\sim$ 95% of all Yeast proteins carry only one type of domain the construction of a domain interaction graph, G_{d-d}, focuses on interactions involving these particular proteins. Thus, the set of nodes consists of domains which appear in interactions of single sorted domain proteins. Obviously, ambiguity arising from multi-domain interactions is thus avoided. An undirected edge between these domains indicates this relationship.

4.2.2 Sources of protein-protein interaction data

Sets of Yeast protein-protein interactions were collected from several overlapping data compilations (Ito et al. 2000; Uetz et al. 2000; Ito et al. 2001; Schwikowski et al. 2000) which employed Yeast two-hybrid experiments extensively. Other relevant interaction data was retrieved from several other protein

interaction databases. The database of interacting proteins (DIP, http://dip.doe-mbi.ucla.edu) scans the literature in order to provide a collection of all functional linkages of proteins obtained by experimental methods (Xenarios et al. 2001). The MIPS Yeast Genome Database (MYGD, http://www.mips.biochem.mpg.de/) (Mewes et al. 2000) is a collection of genetic data from literature which relies on the results of micro array expression experiments. Additionally, MIPS also contains data from Yeast two-hybrid and co-immunoprecipitation experiments.

The protein nomenclature of these data is inconsistent, therefore, the terms were translated to Swiss-Prot/TrEMBL annotations.

4.2.3 Proteome specific data

Yeast specific proteome and protein domain information came from the InterPro database (http://www.ebi.ac.uk/interpro) (Apweiler et al. 2001a) and the Proteome Analysis database (http://www.ebi.ac.uk/proteome) (Apweiler et al. 2001b). Since InterPro employs Swiss-Prot annotation every protein sequence is itemized with each of its domains.

4.2.4 Network properties

From a theoretical point of view, network topologies set up very differently as was already shown. From the results of chapter 3, it is conjectured that all networks considered show a considerable amount of small-world and scale-free characteristics.

In order to unravel the topology of an otherwise unknown network, different characteristic values have been defined. In the networks the degree k_i of a node i is the number of other nodes to which it is connected.

The mean path length of a node i, L_i, is defined as the average of all shortest paths from node i to all other nodes. Accordingly, the mean path length L of the whole network is represented as the average of L_i over all i.

The clustering coefficient of a node i, C_i, measures the fraction of nodes connected to i which are also connected to each other. By extension, the clustering coefficient C of the graph is defined as the average of C_i over all i.

Provided that there exists a sequence of edges $a - b - c$, one might ask to which extent edge $a - c$ are undirectedly linked in the graph. The transitivity coefficient, T_i, represents the mean fraction of neighboring nodes of i which obey this relation. Accordingly, the mean transitivity coefficient, T, is defined as the average of all T_i over all v.

4.2.5 Lethal and viable proteins

Information about lethality and viability of proteins was retrieved from the YPD database (http://www.proteome.com) (Costanzo et al. 2001). Obviously, if one protein proves to be lethal all links in the protein-protein interaction network have to be considered as lethally affected. For the domain-related networks, only fractions of connections prove to be lethal or viable depending on the protein under consideration. Since these networks map protein specific information to a domain dependent space, every link between protein domains does not have to be inevitably proven either lethal or viable. If there exists a domain link which occurs both in one lethal and one viable protein fraction of lethal connections turns out to be 0.5. Hence, these nodes and edges are interesting objects to investigate in regards to their influence on network properties.

4.2.6 Domain fusion events

Since the domain interaction graph focuses on interactions of proteins which carry only one type of domain, the superposition of the protein domain sequence network onto the domain interaction network enables detection of all interaction processes which are accompanied by a domain fusion event on the sequence level. The extent to which domain interactions in Yeast coincide with domain

fusions in higher eukaryotes is of particular interest. Species dependent proteome information was retrieved from Proteome Analysis database.

4.2.7 Graph tools

Graph analysis tools were written in C++ using the LEDA library of data types (Mehlhorn and Naeher 1999).

4.3 Results

4.3.1 Network topologies

It is the intention of this work to provide a comparison of these three networks types. Thus, some already known results are addressed partially in order to provide a thorough view.

All networks emerge as sparse networks providing mean numbers of edges per vertex $\langle k_v \rangle$ which are far smaller than the maximal possible degree per vertex as shown in Table 4.1.

	G_D	G_{d-d}	G_{p-p}
n_v	1196	394	3212
$\langle k_v \rangle$	1,49	3,06	3,79
$n_{conn.\,comp.}$	653	19	89

Table 4.1: Some basic data of the domain sequence, G_D, domain interaction, G_{d-d} and protein-protein interaction graphs, G_{p-p}.

Frequency distributions of links immediately reveal the presence of scale-free topology. Thus, frequency distributions follow a power-law $P(k) \approx k^{-\gamma}$ (Jeong et al. 2001; Park et al. 2001; Wuchty 2001). Figure 4.1 compares the frequency plots of the networks considered. As a result of this analysis, the curves

of frequency distributions of InterPro domain interactions and protein-protein interactions almost coincide. Thus, the assumption that $\sim 95\%$ of the interacting proteins carry only one domain is well reflected.

	L/L_r	C/C_r	T/T_r
G_D	5,25/13,33	0,1989/0,0012	$0{,}0401/4 \times 10^{-4}$
G_{d-d}	4,01/5,18	0,0816/0,0031	0,0296/0.0031
G_{p-p}	4,85/6,18	$0{,}0806/4 \times 10^{-4}$	$0.1288/9 \times 10^{-4}$

Table 4.2: Mean path lengths, L, clustering coefficient, C, and transitivity coefficient, T of the domain sequence, G_D, domain interaction, G_{d-d}, protein-protein interaction graphs, G_{p-p}, and respective random graphs ($_r$).

Additionally, InterPro domain sequence networks have been found to exhibit small-world properties (Wuchty 2001). Addressing the relevant parameters of small-worldedness, the clustering coefficient C of InterPro domain sequence networks far exceeds the respective one of an equally sized random graph. However, the definition of small-world networks additionally demands $L \geq L_{random}$. It turns out that the major component of the domain sequence network which covers the majority of domains fulfills this demand. As a result, protein and domain interaction networks both feature clustering coefficients which fulfill the definition of small-world networks (Table 4.2). However, respective numbers of L fail. Although protein interaction and domain interaction networks both feature huge 'major' components neither of them satisfies this structural demand of small-world networks.

4.3.2 Biological hubs

Scale-free topology suggests that only a minority of the highly connected nodes shape the topology of the underlying network. Since highly connected hubs play a crucial role for information processing and integrity of networks, it is interesting to see which role these nodes play in biological networks. Table 4.3.2 shows the 15 most highly connected nodes of each network. Highest linked InterPro

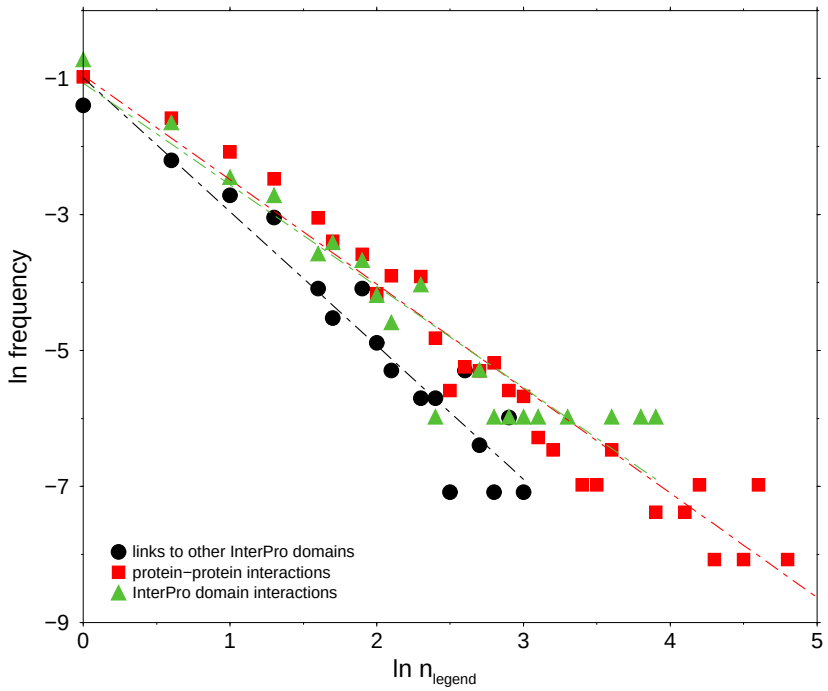


Figure 4.1: The frequency distribution of the protein-protein interaction graph, G_{p-p}, domain interaction graph, G_{d-d} and domain sequence graph, G_D. The numbers of links to other vertices were logarithmically binned and frequencies thus obtained.

domains in domain sequence networks of Yeast were already found to be involved in signal transduction pathways. Other high linked domains appear in transcriptional/translational activities and energy maintenance (Wuchty 2001).

In this analysis, the significance of signalling pathways is strongly emphasised in the domain interaction network by WD40 and zinc-finger motifs which are among the highest interacting domains.

Strongest interacting proteins are involved in nucleus related transportation processes. These include subunits of Importin and nucleoporins. Furthermore, cell-cycle regulating (MEC3, TEM1) and transcription processing proteins appear highly interacting.

$v^{G_{p-p}}$	k_v	$v^{G_{d-d}}$	k_v	v^{G_D}	k_v
JSN1	230	WD40	53	zf-C3HC4	21
Importin α subunit	123	RRM	46	pkinase	19
ATP14	108	Zn2-CY6-fungal	39	Ser-Thr-kin-actsite	19
TIR1 precursor	107	snRNP-Sm	28	AAA	19
NUP116/NSP116	107	vATP-synt AC39	24	PH	18
SRB4	92	zf-C2H2	22	EF-hand	16
TFIIB	81	cyclin	20	C2	15
YHR4	72	Ser/Thr-phosphat.	18	WD40	14
VMA6	71	TPR	16	DEAD	14
PGDH	69	SH3	15	helicase C	14
MEC3	65	bZIP	12	ATP-GTP-A	14
TEM1 protein	61	TFIID	11	AA-tRNA-ligase-II	14
SOH1 protein	50	Myb-DNA-bind	11	CPSase	14
LYS14 protein	50	zf-CCHC	11	GATase-1	13
Importin β-1 sub.	40	Histone-core	11	FMN-binding enz.	12

Table 4.3: The 15 most highly connected nodes of the protein-protein interaction graph, G_{p-p}, the domain interaction graph, G_{d-d}, and the domain sequence graph, G_D.

4.3.3 Transitivity

Since scale-free and small-world networks were found to be sparse but highly clustered, the degree of transitivity can be questioned. The mean transitivity value, T, measures the extent to which indirect links are accompanied by direct ones. Such a 'back-up' of links reinforces the clustered nature of biological (sub)networks. Table 4.2 shows statistics of the networks under this consideration. Similarly to the behavior of C, T exceeds the respective number of random graphs of equal size, T_r, by far. However, it should be noted that the values are reasonably low.

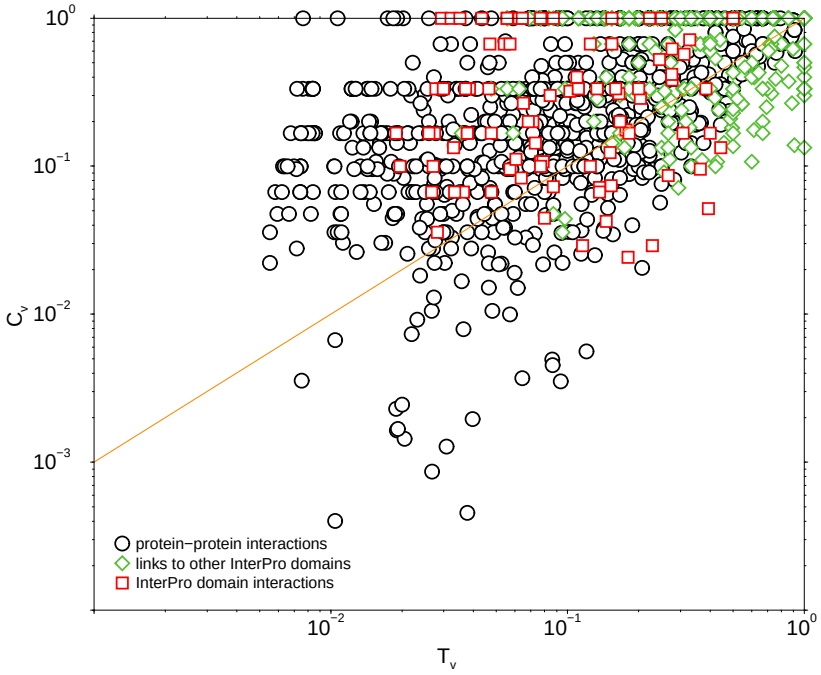


Figure 4.2: Scatterplot of mean clustering coefficient C_v vs. mean transitivity coefficient C_v concerning the protein-protein interaction graph, G_{p-p}, domain interaction graph, G_{d-d}, and domain sequence graph, G_D.

It might be tempting to assume that T is closely related to C since an edge $a-c$ implies an increase of $C(b)$. In order to investigate the mutual relation of T and C Figure 4.2 shows a scatterplot of T against C concerning all three types of networks. Considering Figure 4.2, symbols indeed arrange around the median axis. However, they are far from indicating a strong correlation.

From a biological point of view, it is interesting to discover the role of proteins which are involved in such a transitive organisation. Table 4.3.3 shows a compilation of proteins and domains exhibiting highest T_v. Strikingly, the list of interacting proteins is headed by proteins which form enzymatic protein clusters. Among them are PMT and OST proteins setting up Dolichyl-Diphosphooligosaccharide-protein glycosyltransferase protein complex. Similarly, the interacting domains with the highest T-values are protagonists of functional clusters involved in transcription (TFIID-proteins and RNA pol β subunit) and signal transduction. However, the T-values of interacting domains are lower than those of interacting

$v^{G_{p-p}}$	T_v	$v^{G_{d-d}}$	T_v	v^{G_D}	T_v
WBP1	0,92	STT3	0,50	DNAtopI-DNA-bind	1,0
PMT3	0,91	DAD	0,50	DNAtopI-ATP-bind	1,0
PMT4	0,91	RNA pol β s.u.	0.44	DNA pol β-like	1,0
PMT2	0,91	MCM	0,40	Interleukin-1	1,0
OST5	0,90	WD40	0,39	RNA-polII-repeat	1,0
OST3	0,90	T-SNARE	0,38	ATPase-α-β	1,0
OST2	0,90	Ribosomal-S12	0.36	Tubulin	1,0
STT3	0,90	BK-channel-α	0,33	CytC-heme-bind	1,0
SWP1	0,89	TFIID-31	0,31	6-P-fructo-2-kinase	1,0
UBC5	0,80	Synaptobrevin	0,31	RNA-pol-A	1,0
UBC4	0,80	TFIID-18	0,28	Gluc-transporter	1,0
OST4	0,79	Znf-CCHC	0,28	Dynamin	1,0
ALG5	0,76	Histone-core	0,28	Helix-hairp.-helix motif cl. 2	1,0
VPS16	0,75	Znf-C2H2	0,27	Middle domain of eIF4G	1,0
PEP3	0,75	DNA-RNApol-7kD	0,25	GHMP-kinase	1,0

Table 4.4: The 15 most transitive nodes of the protein-protein interaction network, G_{p-p}, domain interaction network, G_{d-d}, and domain sequence network, G_D.

proteins. The picture changes drastically if T-values of the domain sequence network are considered since these tend to be shifted to higher values.

4.3.4 Lethality and Viability

The separation of viable and lethal proteins allows one to observe protein interaction networks and domain related networks from a different perspective. The frequency distributions of both lethal and viable proteins in the protein-protein interaction network are shown in Figure 4.3.

Initially, it was suggested that strongly interacting proteins can be assigned a

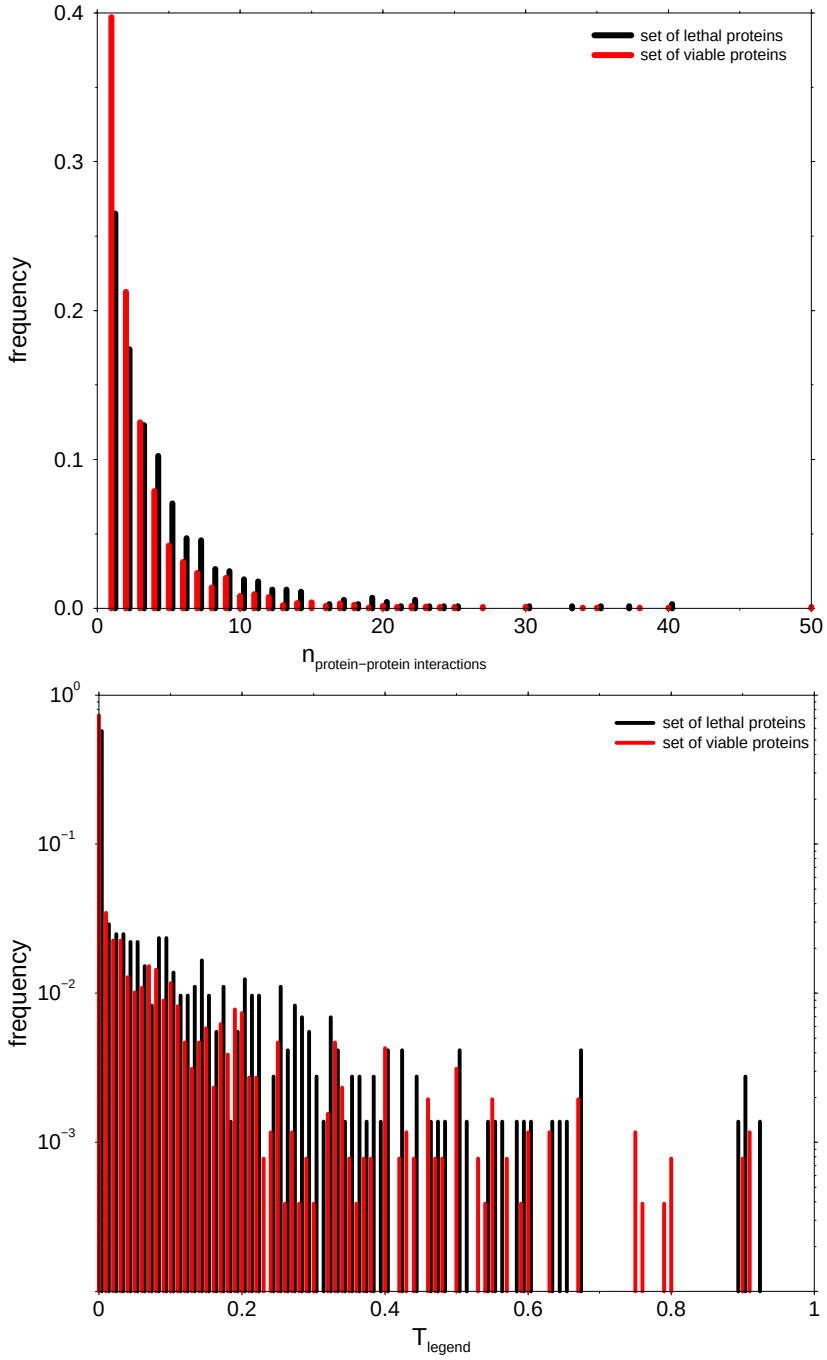


Figure 4.3: Frequency distributions of the degree, k (upper panel), and mean transitivity coefficient, T_v (lower panel), which are set up by interactions of lethal and viable proteins. The compilation of lethal and viable proteins was retrieved from the YPD database.

lethal role (Jeong et al. 2001). In fact, it appears in this analysis that this assumption is misleading since the latter plot indicates merely a slight trend of lethal proteins to accumulate higher numbers of interactions than viable ones. Since the transitivity coefficient takes the existence of alternative paths into account, it might be interesting to check if T is more suited to explain the latter correlation. Figure 4.3 shows a frequency plot of T regarding lethal and viable proteins. Confirming the latter assumption lethal proteins indicate a slight trend to higher T. Regarding higher values of T, it clearly appears that lethal proteins tend to accumulate more alternative interaction paths. However, it should be noted that frequencies are considerably low. A similar view holds for lethal and viable fractions of domains in the respective networks. Figure 4.4 displays frequency distributions of fractions of lethal and viable domain interactions and domain connections in the respective graphs. Analogously, the plots suggest a slight shift to lower fractions of lethal connections in Figure 4.4.

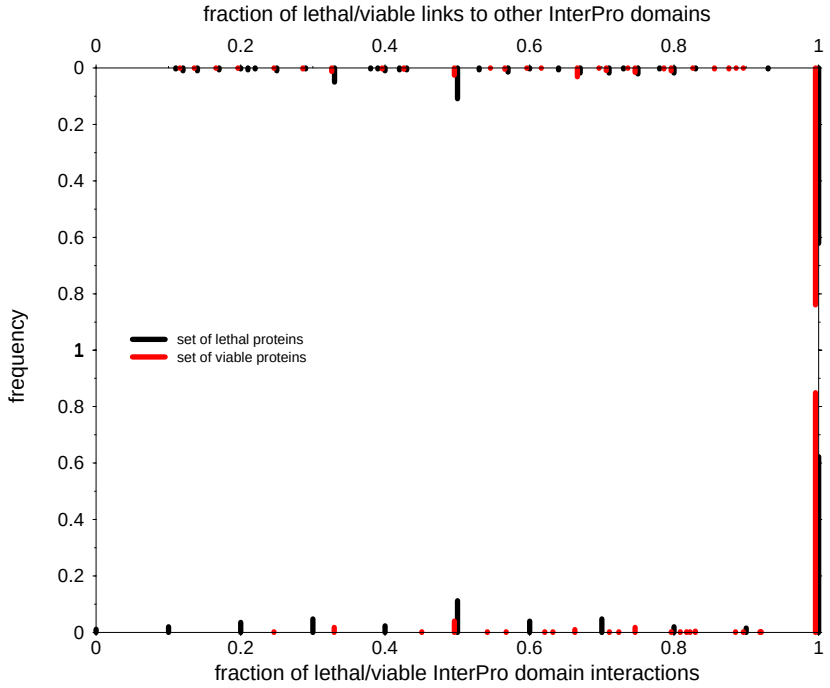


Figure 4.4: Frequency distributions of fraction of lethal and viable links per domain in domain interaction and domain sequence networks. The compilation of lethal and viable proteins was retrieved from the YPD database.

Observing the mutational effects from a different perspective, Figure 4.5 displays frequency distributions of the mean path lengths L and mean clustering coefficients C of the protein-protein interaction network. Results were obtained by deleting separately lethally and viably mutated proteins and subsequent calculation of these network properties. Both types of distributions are normally distributed. The distributions of lethally perturbed networks generally show slightly increased standard deviations. These observations also hold for domain related networks which were considered analogously by clipping fractions of links affected by lethal or viable mutations of the respective proteins.

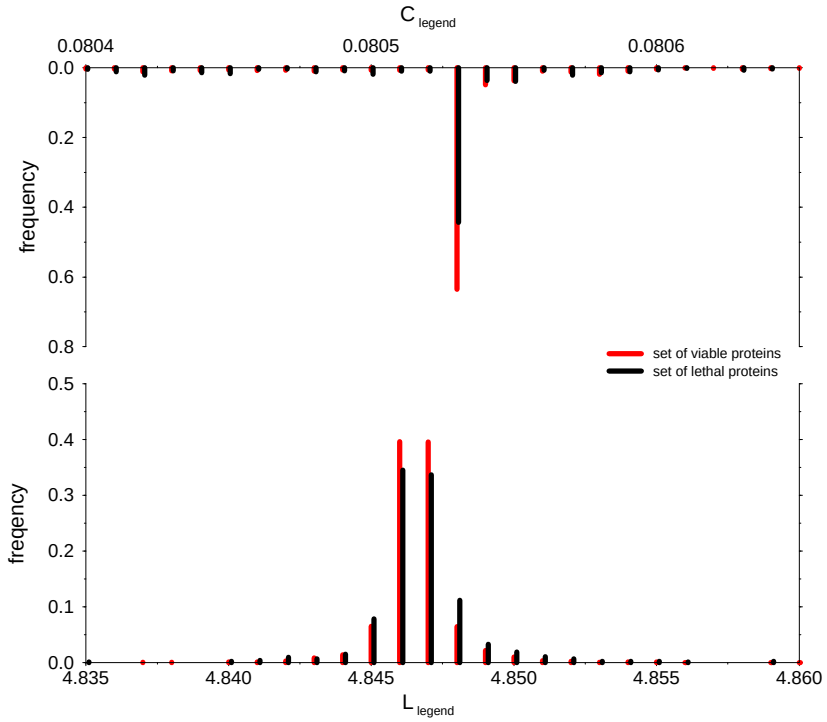


Figure 4.5: Distributions of the mean path length L_v and mean clustering coefficient C_v of the protein-protein interaction network. Lethally and viably mutated proteins were clipped and network parameters thus obtained. Protein information was retrieved from the YPD database.

4.3.5 Domain interactions and fusion events

Functional links between proteins have also been detected by analyzing fusion patterns of protein domains. Separate proteins A and B in one organism are

found to be expressed as a fusion protein in other species. A protein sequence containing both A and B is termed a Rosetta Stone sequence (Marcotte et al. 1999). The comparison of pairwise domain interactions and pairwise domain fusions in higher organisms enables an estimation of the extent to which domain interactions are indeed accompanied by a domain fusion event. Pairs of domain interactions and domain links correspond to edges in the respective networks. Considering every domain separately edges in the Yeast domain interaction network are counted which co-occur in the domain sequence networks of *A. thaliana*, *C. elegans*, *Drosophila*, *H.sapiens* and Yeast, respectively. Subsequently, fractions of domain fusion per domain interaction of the mentioned organisms are calculated. Figure 4.6 summarizes as a result an increasing extent of domain fusions in the latter row of eukaryotes.

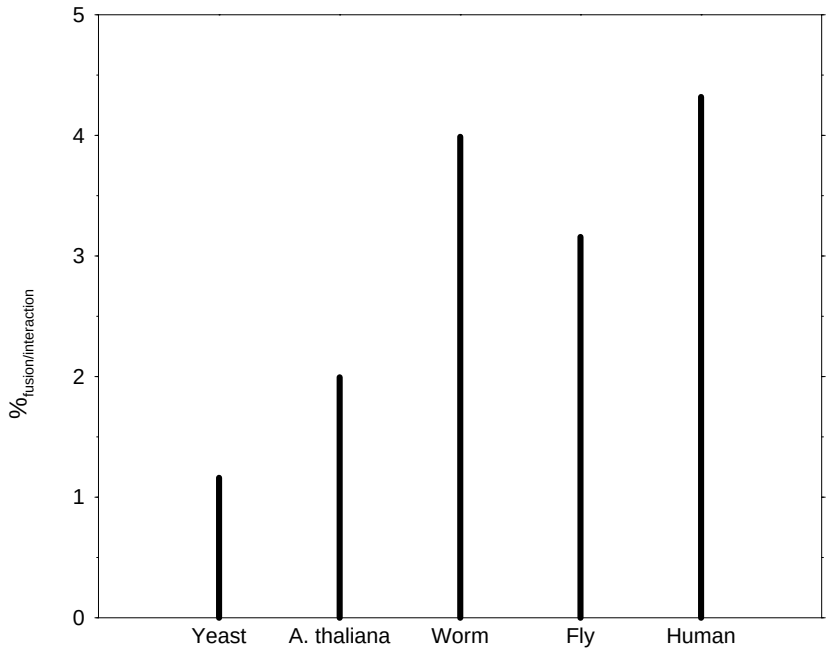


Figure 4.6: Histogram of domain fusion events per domain interaction. The co-occurrence of domain interactions found in *S.cerevisiae* and domain fusion events were detected in *S.cerevisiae*, *A. thaliana*, *C.elegans*, *Drosophila* and *H.sapiens*. Domain information was obtained from InterPro domain database.

4.4 Discussion

4.4.1 Completeness and quality of data

The protein-protein interaction data used for the set up of the interaction network are widely based on yeast two-hybrid analyses. However, yeast two-hybrid data are significantly flawed by high rates of false positive signals (Hazbun and Fields 2001). Moreover, many of the interactions identified merely rely on positive signals from one single technique and result from indirect observations. The observation that Importin α subunit protein (SRP1) (Table 4.3.2) interacts with that number of proteins is merely a result of the two-hybrid screen employed since a very small fraction of those interactions were shown by other methods.

The discovery of scale-freedom in protein and domain related networks alleviates the insurmountable problems arising from the current extent of incompleteness. Even though the current interaction data are far from complete and are somewhat noisy, these findings reinforce the argument that the topology of interaction networks will not change significantly as the amount of interaction data grows.

Strictly speaking, the set up of the domain interaction network is an indirect one since interactions are inferred from protein interactions and domain sequence information. In contrast to other approaches, no structural information of domains was taken into account. Thus, it should be noted that domain interaction networks mediate a certain degree of simplification. However, even though the domain interaction network is simplified to a certain degree scale-freeness of the network confirms the assumption that the topology will not change with increasing amount and quality of domain and protein interaction data.

Analogously, this assumption also holds for domain sequence networks since the proteome data have been far better compiled and studied with the release of the complete genomic sequence of *Saccharomyces cerevisiae*. The assumption of scale-free characteristics leads to interaction and domain sequence graphs independent of the actual size of the underlying networks. Although the generation

and compilation of interaction data is still at a basic level these interaction graphs give tentative insights of the underlying network topology.

4.4.2 What do these network architectures tell?

The observation of scale-freedom in all three networks confirms the appearance of sparse but highly clustered nets. As a consequence, highly connected nodes emerge which predominantly shape the topology of the underlying network. Considering the shortest ways through the network, it will become immediately clear that these routes always pass highly connected nodes. Thus, these hubs illustrate crossways helping to transport information quickly to even remote parts of the network.

So, sparsity and strong local clustering of the scale-free nets offer a different view on the organization of the networks considered. Pathways defined by protein and domain interactions might be treated as highly clustered subnets which are sparsely interlinked to other ones. Accordingly, highly interacting proteins and domains can be considered as the 'backbone' of the networks which interconnect pathways in the respective networks. Otherwise, these nodes might be central proteins and domains which shape a particular pathway. Thus, it is possible to get a good flavor of the general characteristics of the underlying networks without the knowledge of all interactions. This idea intuitively becomes important since the current interaction data are from being complete as already mentioned in the latter section.

In order to get a flavor how frequent sequences of edges $a - b - c$ are accompanied by co-occuring edges $a - c$, a new measure, mean transitivity coefficient T, was introduced. Similarly to mean clustering coefficient C, T-values of scale-free and small-world networks exceed the respective numbers of equally sized random graphs strongly emphasizing a tendency to reinforce clustering. However, the T-values of all three networks of Yeast indicate the assumption that indirect or alternative linkage might be rather an exceptional than common feature. However, the latter point crucially depends on the set up of networks. Since domain

networks were generated by considering domain nodes linked if they co-occur with other ones in proteins, T is subject to a shift to higher values. Nevertheless, this value reflects the extent to which particular sequences of nodes are 'backed up' by an inserted direct link. As already mentioned, proteins which are mainly involved in enzymatic clusters display a high degree of transitivity. Obviously, this result is based on the observation that these proteins nearly interact with each other in the respective protein clusters. Otherwise, two nodes - although already linked - might be connected indirectly by adding an intermediate node. Considering protein and domain interactions, intermediate proteins and domains might be considered as the entry to alternative pathways. Analogously, intermediate domains in domain sequence networks might display access to different domain architectures. Thus, high values of T imply domains which frequently co-occur with the same domains. Since a crucial role for the networks topology coincides with connectedness, highly transitive nodes might be among the sets of highly connected nodes of the networks considered. However, no evidence to support this assumption was found. In fact, it turns out that rather the opposite is the case since frequent interacting proteins like JSN1 or YHR4 show transitivity coefficients around 0.03. The same holds for highest connected domains in the domain sequence network emphasizing pkinase and WD40 representing transitivity coefficients of 0.49 and 0.09, respectively. In contrast, domains of the domain interaction network apparently contradict this particular trend since highly interacting domains show reasonable high transitivity more frequently. WD40 and RRM which lead the list of highly interacting domains in Table 4.3.2 emerge as fairly transitive with values of 0.39 and 0.29, respectively. However, this apparent contradictory trend seems to be more the result of the small sample than a characteristic of interacting domains.

4.4.3 Evolutionary aspects

Compared to the Yeast proteome, domain fusions in the proteomes of higher organism are more frequent (Marcotte et al. 1999). On the one hand, proteome complexity is particularly assumed to be the consequence of protein innovations. On the other hand, proteomes are generated by expansion of protein families

and subsequent combinatorial arrangements of domains. Combinatorial diversity provides protein sets which are sufficient to preserve cellular procedures without dramatically expanding the absolute size of the protein complement. The list of highly connected domains in domain sequence networks immediately reveals a substantial lack of overlap in the compilation of single interacting domains. Although the ratios of fusions grow constantly towards organisms of increasing complexity they remain considerably low. Subsequent fusion of interacting domains seems to be rather an exceptional than a common feature. Accordingly, single domain interactions seem to be no driving force for fusing domains in one sequence. Naturally, one might argue that the number of fusions will be subject to tremendous change when the Yeast interactome will be further explored. Since the knowledge about the Yeast interactome is far from being complete, the overall trend of increasing numbers of fusions per domain interaction will still be reflected by improved numbers.

Considering highly connected domain nodes in Table 4.3.2, the abundance of domains involved in signal transduction pathways like kinases and zinc-finger motifs is conspicuous. Proteins which emerged by fusion of domains or combinatorial diversification of domain architectures are important parts of signal transduction and cell-cell communication pathways of Yeast emphasizing its role as a single cellular organism leading the way to multicellularity. Domains which proved to be fit in different cellular aspects of Yeast are rewarded with an increasing degree of connections in higher eukaryotes emphasizing a sort of 'fit-get-rich' regime (Wuchty 2001). Thus, it might be expectable that these partially highly connected proteins and domains identify as very crucial for the survivability of the cell. However, it turned out that this is not the case. Although perturbation analysis of all three types of networks indicates a tendency of lethal proteins and domains to slightly assemble more crucial effects on the networks the results are far from offering a clear distinction between lethal and viable sets of proteins and domains. However, it should be kept in mind that this results might be based on the low complexity of Yeast and absence of highly comprehensive data sets. With protein specific data of higher organism this question will be revisited.

CHAPTER 5

The large scale organization of genomic sequence segments

5.1 Introduction

The abundance of fully sequenced genomes of different organisms inspired researchers to ask for genomic homogeneity and heterogeneity in terms of multinucleotide relative abundances and compositional extremes. In the recent years, many contributions to these questions have been published. Henceforth, the current presence of completely sequenced eukaryotic genomes adds additional weight to the impact of such investigations. Comparative studies have mostly focused on short oligonucleotides such as dinucleotides (Burge et al. 1992; Karlin and Ladunga 1994; Karlin and Burge 1995; Blaisdell et al. 1996; Karlin and Mrázek 1997; Karlin et al. 1997; Nakashima et al. 1998), trinucleotides (Burge et al. 1992; Karlin and Ladunga 1994; Karlin et al. 1994; Karlin and Mrázek 1997; Karlin et al. 1997) and tetranucleotides (Karlin and Ladunga 1994; Karlin et al. 1997). Differences of motifs up to eight nucleotides were investigated using chaos game representation (Deschavanne et al. 1999). The tremendous amount of results supported the looming pattern that intergenomic differences are higher than intragenomic ones (Karlin and Ladunga 1994; Karlin et al. 1994; Blaisdell et al. 1996; Karlin and Mrázek 1997; Karlin et al. 1997; Gentles and Karlin

2001; Nakashima et al. 1998; Deschavanne et al. 1999). The abundance and constantly varying frequencies of oligonucleotides at different sites in genomic sequences inspired researchers to introduce the notion of characteristic dinucleotide genomic signatures of every organism (Karlin and Ladunga 1994; Gentles and Karlin 2001). Essentially, it is constant in both coding and noncoding sequences and does not depend on knowledge of individual genes. Furthermore, the existence of specific genomic signatures for all motif lengths has been indicated (Deschavanne et al. 1999).

Beyond the biological scope, there have also been a variety of works which addressed statistical implications of genomic sequence composition. One of the pioneering works suggested introns to show long-range correlations in contrast to exons (Peng et al. 1992). The presence of these correlations in introns and noncoding sequences nourished the presumption that these sequences rather feature linguistic properties than exons or coding sequences (Mantegna et al. 1994; Mantegna et al. 1995; Martindale and Konopka 1996). Henceforth, genomic sequences were investigated emphasizing linguistic methods (Mantegna et al. 1995; Stanley et al. 1999) leading to the emergence of coding potentials of genomic sequences (Grosse et al. 2000; Holste et al. 2000; Fickett and Tung 1992).

Most recently, a method was suggested to classify the genomic origin of bacterial sequences. Based on relative frequencies of oligonucleotides, a Bayesian classifier was set up which - interestingly enough - enhances its predictive power with increasing length of oligonucleotides (Sandberg et al. 2001).

However, one might have the impression that the works which consider properties of segmentwise composition of DNA provide rather patchwork character than a compact view of DNA features. Although there have been so many investigations up to the present time a more common perspective is still lacking which considers the features segmentations of different lengths mediate.

Obviously, the length of sequences opens just as many opportunities to start the segmentation if shifts of one nucleotide are considered. Essentially, each seg-

mentation process of length l results in l sets of overlapping segments. Since there exists a distinct $5' \rightarrow 3'$ reading direction on genomic sequences, immediate subsequent segments in each of the underlying l segments sets are regarded as adjacent. This approach of segments resembles the set up of a graph. Thus, a genetic segments graph, $G_S = (V_S, E_S)$, is formally defined by a nodes set, V_S, consisting of all segments of length l found within a set of genomic sequences. Since two segments are found to be subsequently adjacent in $5' \rightarrow 3'$ direction at least one time, a new directed link is added to the set of links, E_S. In a directed network, the degree k_{in} of a node is the numbers of other nodes to which it points. The same notion applies for the degree k_{out} denoting the number of nodes which themselves point to this particular one.

In the following, I will present results which treat segmental composition of exons and intron sequences of different organisms from a networks standpoint. Results thus emerging will be discussed from the perspective of already mentioned findings. Finally, classification power of exon and intron segments segmented to different lengths will be addressed.

5.2 Materials and methods

5.2.1 Genomic sequence data

As a source of well curated samples of eukaryotic exon and intron sequences, the Exon-Intron-Database (EID, http://golgi.harvard.edu/gilbert/eid/) was chosen which is based on GenBank 115 entries. EID provides exhaustive data about eukaryotic protein-coding and intron-containing genes (Saxonov et al. 2000). In order to emphasize differences which separates natural occurring sequences clearly from random ones, random pendants of natural sequence sets were generated.

5.2.2 Network density

An inverse measure of sparseness, the network density (i Cancho and Solé 2001), is defined as

$$\rho = \frac{E}{N^2}, \tag{5.1}$$

where E is the number of edges and N is the number of nodes.

5.2.3 Measures of divergence

The relative entropy, also known as Kulback-Leibler divergence, is a measure how different two probability distributions over the same event space are (Manning and Schütze 1999). Consider

$$P(k_i^{out}) = \frac{k_i^{out}}{\sum_j k_j^{out}} \tag{5.2}$$

as the frequency of the number of links pointing from node i over the total number of outgoing links in the underlying graph. Analogously, a distribution $P_r(k_i^{out})$ can also be obtained for the respective node in a genetic segment graph which was set up with random sequences of the same sample size. Thus, the relative entropy,

$$H(P \parallel P_r) = \sum_i P(k_i^{out}) \log_2 \frac{P(k_i^{out})}{P_r(k_i^{out})} \tag{5.3}$$

measures the distance between the natural and random distribution of genomic segments.

5.2.4 Classification of segments

Segments of different length raise the question if there exists a common distinction of exon and intron sequences at this level. So to speak, are there segments which can more or less clearly classified to an exonic or intronic origin?
Thus, there is the interest in the *a posteriori* probability that the exon model applies for a certain segment, $P(E\,|s)$. This treatment is inspired by Bayesian statistics. Thus,

$$P(E\,|s) = \frac{P(s\,|E)P(E)}{P(s\,|E)P(E) + P(s\,|\,I)(1 - P(E))}, \tag{5.4}$$

applies where s is a segment of particular length. E and I are the samples of segments from exon and intron sequences which specify $P(E)$ and $P(I) = 1 - P(E)$ being the frequencies of all segments in the respective samples. Obviously, this terms reflect *a priori* expectations that segments are related to exon and intron sets before they are actually seen. Henceforth, $P(s\,|E)$ and $P(s\,|\,I)$ denote the frequencies that a particular segment s occurs in the set of exon or intron sequences (Durbin et al. 1998).

5.3 Results

5.3.1 Connectivity distributions

Considering sequences from the perspective of networks which are set up by their subsequent segments of variable length, raises the hope to get new insights in the different composition of exon and intron sequences. Networks thus set up exhibit different distributions of outgoing links depending on the segments size. Intriguingly, networks of exon segments of Yeast display a transition from an Gaussian distribution via a truncated power-law to a real power-law shaped connectivity distribution towards increasing segment size (Figure 5.1).

The same procedure was applied to a set of equally sized set of random sequences. Even more interestingly, the same sequence of transitions is observed albeit explicit differences in the distributions parameters are visible. Qualitatively, the same finding also holds for networks of intron segments of Yeast and their equal sized random pendant (Figure 5.2). However, it is conspicuous that especially connectivity distributions of small sized natural segments are more scattered than the respective random ones.

Coincidentally, this transition is accompanied by a rapid decrease of network densities towards increasing size of segments as Figure 5.3 shows.
Essentially, what can be observed and generalized easily to segment networks of other organisms is the transition from a Gaussian distribution of a dense graph to a sparsely connected scale-free network which is characterized by a real power-law

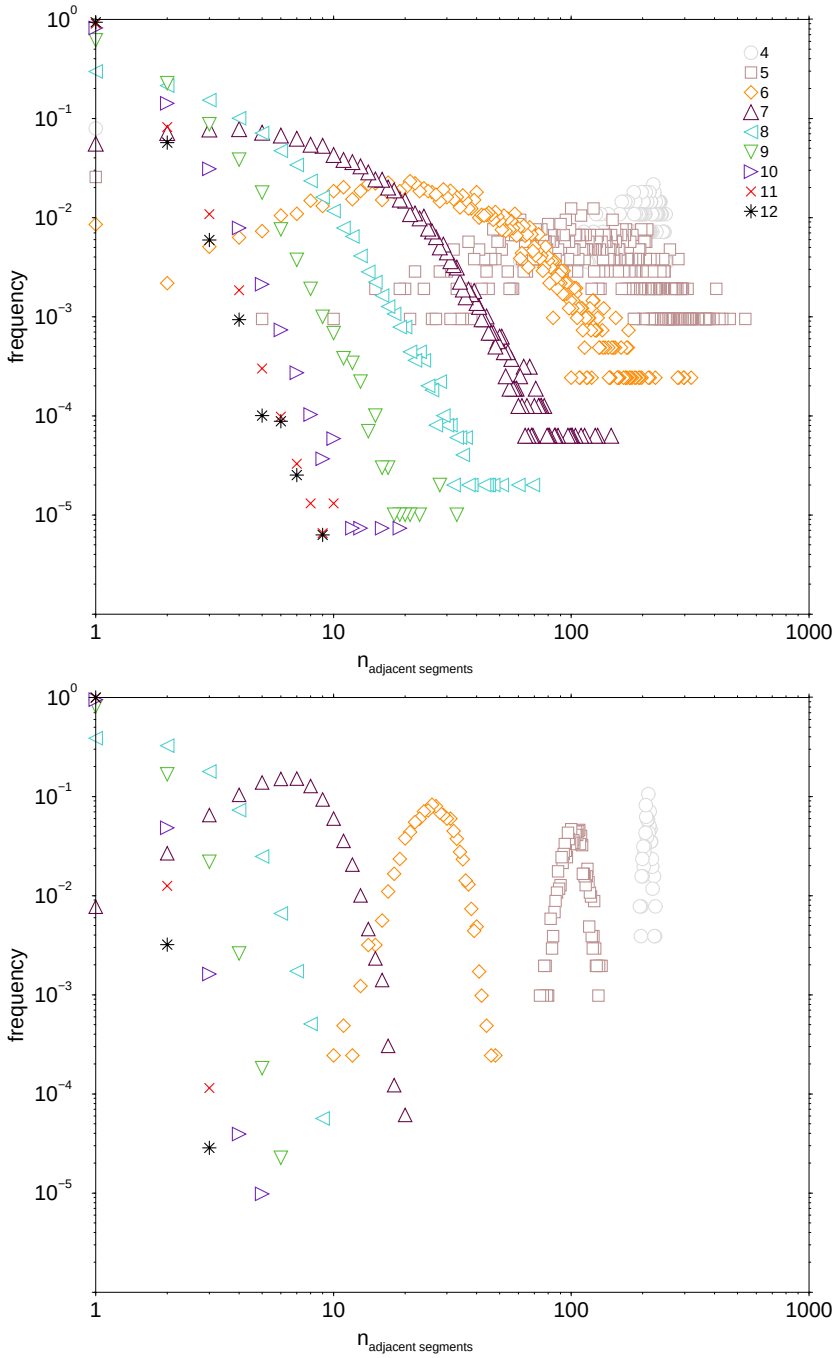


Figure 5.1: Connectivity distributions of segments networks of Yeast exons (upper panel) and a randomly sampled pendant (lower panel). Natural sequence data were retrieved from EID database. Networks are based on segments sizes which range from 4 to 12.

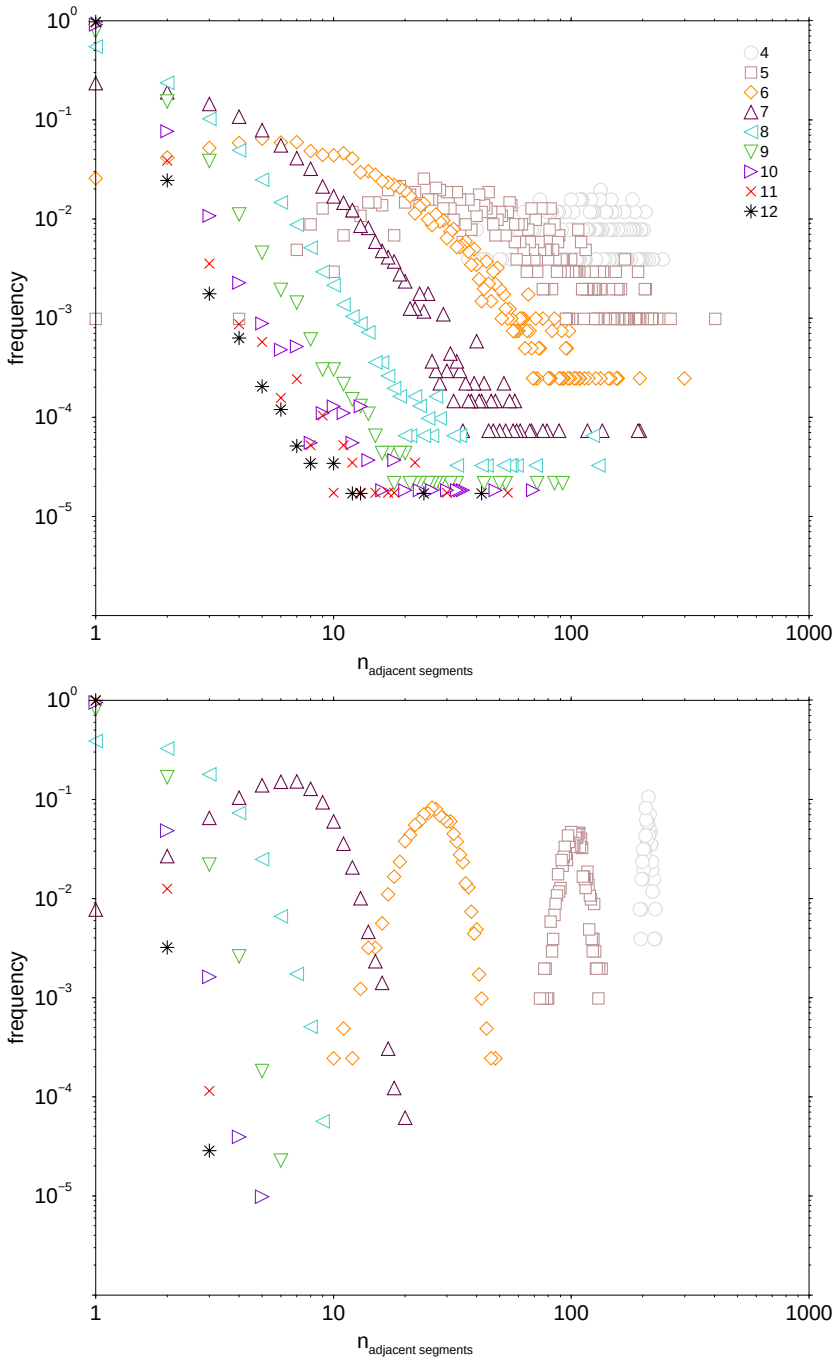


Figure 5.2: Connectivity distributions of segments networks of Yeast introns (upper panel) and a randomly sampled pendant (lower panel). Natural sequence data were retrieved from EID database. Networks are based on segments sizes which range from 4 to 12.

towards increasing size of segments. Since the cross-over to completely oppositional network shapes is clearly not sharp, the intermediate level is represented by a truncated power-law.

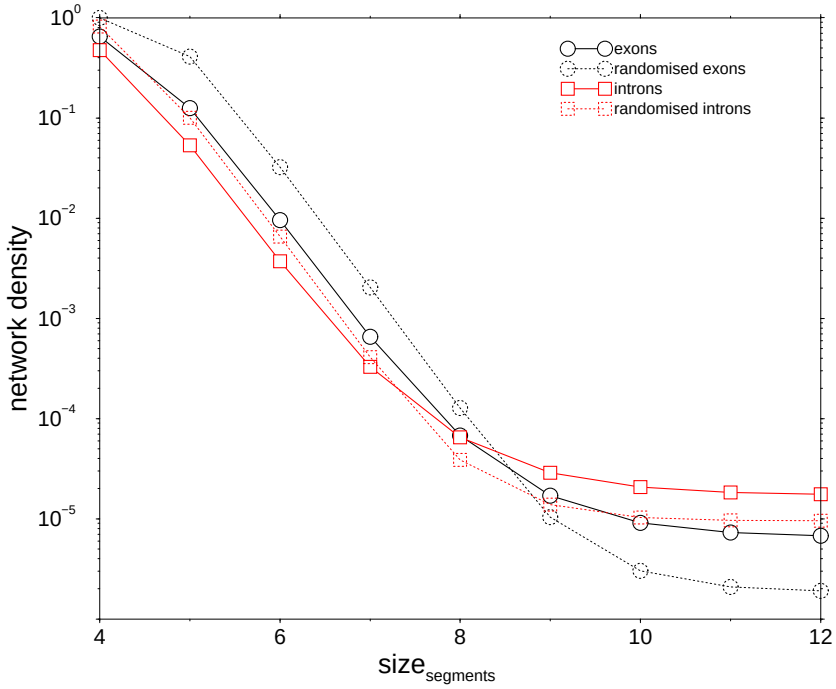


Figure 5.3: Network density distributions of segments networks of Yeast exons, introns and randomly sampled pendants. Natural sequence data were retrieved from EID database. Networks are based on segments sizes which range from 4 to 12.

5.3.2 Divergence of segment networks

Relative entropy is a measure of divergence of two distributions over the same event space. Connectivity distributions of segments networks combined with the respective ones of randomized exon and intron sequences represent such comparable event spaces. Since sets of intron sequences especially of higher eukaryotes have tremendous sizes, the analysis is restricted to a representative random sample. Hence, 10 % of each complete sequence sets size were randomly picked, segmented and networks thus generated. The respective randomized samples of equal size underwent the same procedures of segmentation and network generation. Although networks which were set up by natural and randomized sequences

comprise transitions from random to scale-free network topology with increasing segments size Figure 5.4 unveils reasonable differences with a certain set of segment sizes. Strikingly, these particular segment sizes group around those sets which indicate topology transition.

5.3.3 Classification of exons and intron segments

Segments of different length raise the question if segments can more or less clearly classified to an exonic or intronic origin. The analysis is carried out on the natural and random sequence sets of Yeast and Human. In order to keep computational demands in a reasonable frame but to preserve natural occurring relationships of sample sizes, the same sets of randomly picked exon and intron sequences as well as their randomly generated pendants were chosen for this analysis. Figure 5.5 shows histograms of probability frequencies that segments of different size belong to sets of exon sequences and random exon sequences, respectively. The set size of Yeast exons exceeds the respective intron one while sizes of the human sequence sets behave the other way round. These size effect are detectable with small segment sizes but diminish at latest with segment size 12. Essentially, this observation also holds for the respective random sets of sequences. Regarding segment sizes, histograms clarify that sets of random sequences approach increasing quality of classifications 'faster'.

5.3.4 Connectivity distributions of different eukaryotic organisms

In a recent paper, weak correlations between nucleotides 10-11 units apart were reported (Herzel et al. 1998; Herzel et al. 1999). Size 11 was chosen as segment length to investigate networks of exon and intron sequences of different eukaryotic organisms. Again the sets of randomly picked sequences was used for this analysis. Figure 5.6 shows the results. Both distributions show different slopes for different organisms depending on the organisms complexity. Interestingly, this finding holds for exons as well as for introns although introns show significantly much more irregularities.

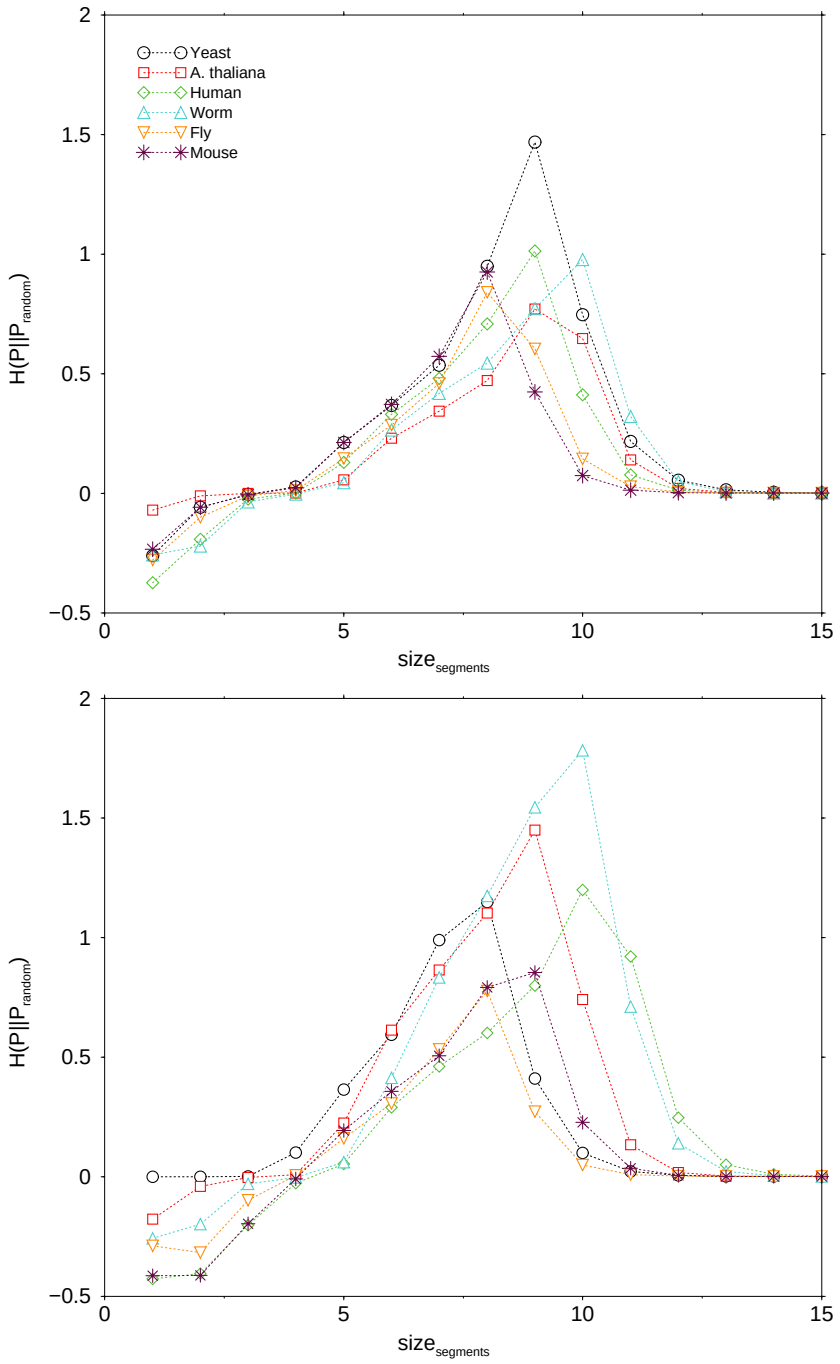


Figure 5.4: Divergence of 'natural' networks set up by exon (upper panel) and intron sequence segments (lower panel) of Human, Fly, Worm, Mouse, Yeast and *Arabidopsis thaliana* and randomly generated pendants. As a representative sample 10 % of each complete sequence sets were randomly picked from EID database. Segment sizes range from 1 to 15.

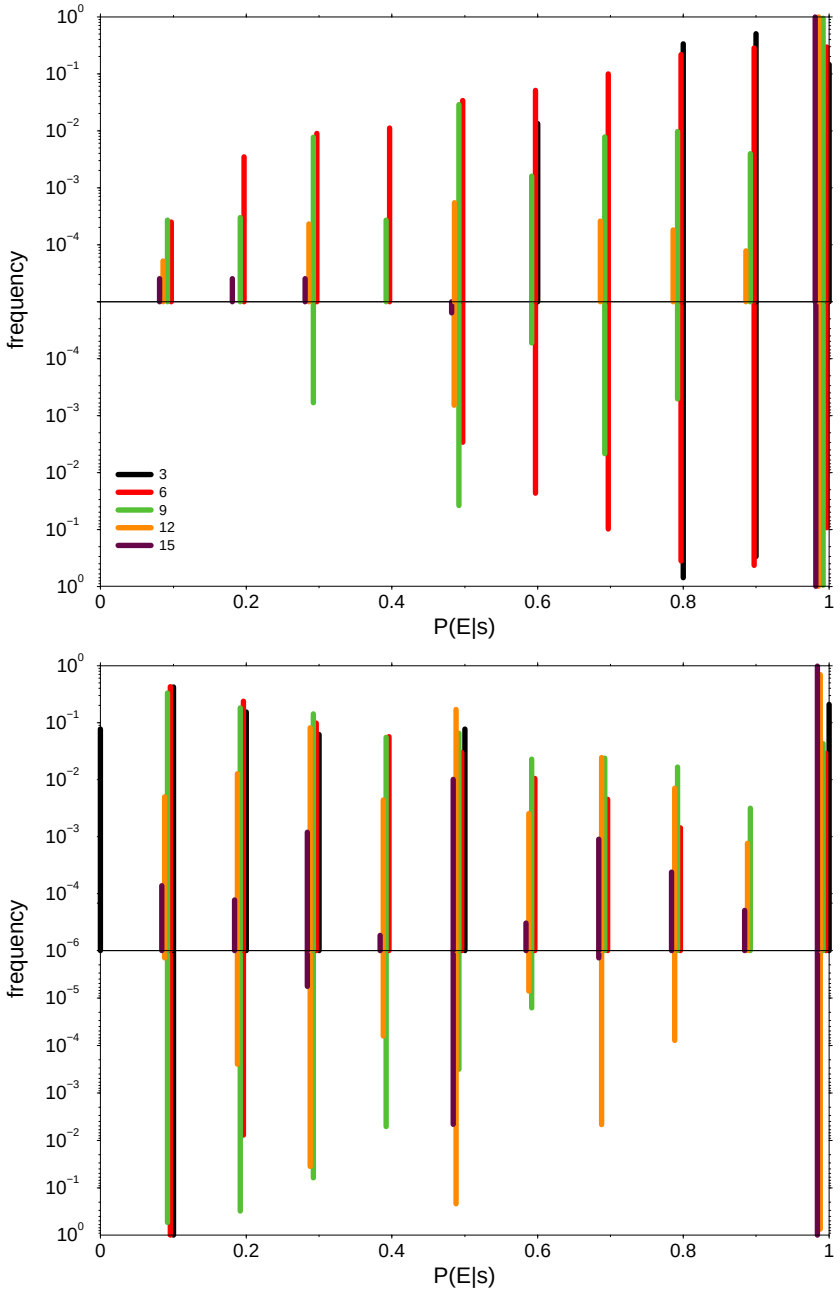


Figure 5.5: Histograms of probabilities that a segment of a certain size belongs to natural or randomized exon sequences of Yeast (upper panel) and Human (lower panel). The upper panels of the diagrams refer to the natural sets of exon sequences. The lower panels correspond to randomized sequences. 10 % of each complete exon and intron sets were randomly picked and randomized sequences thus generated. Exon and intron sets were retrieved from EID database.

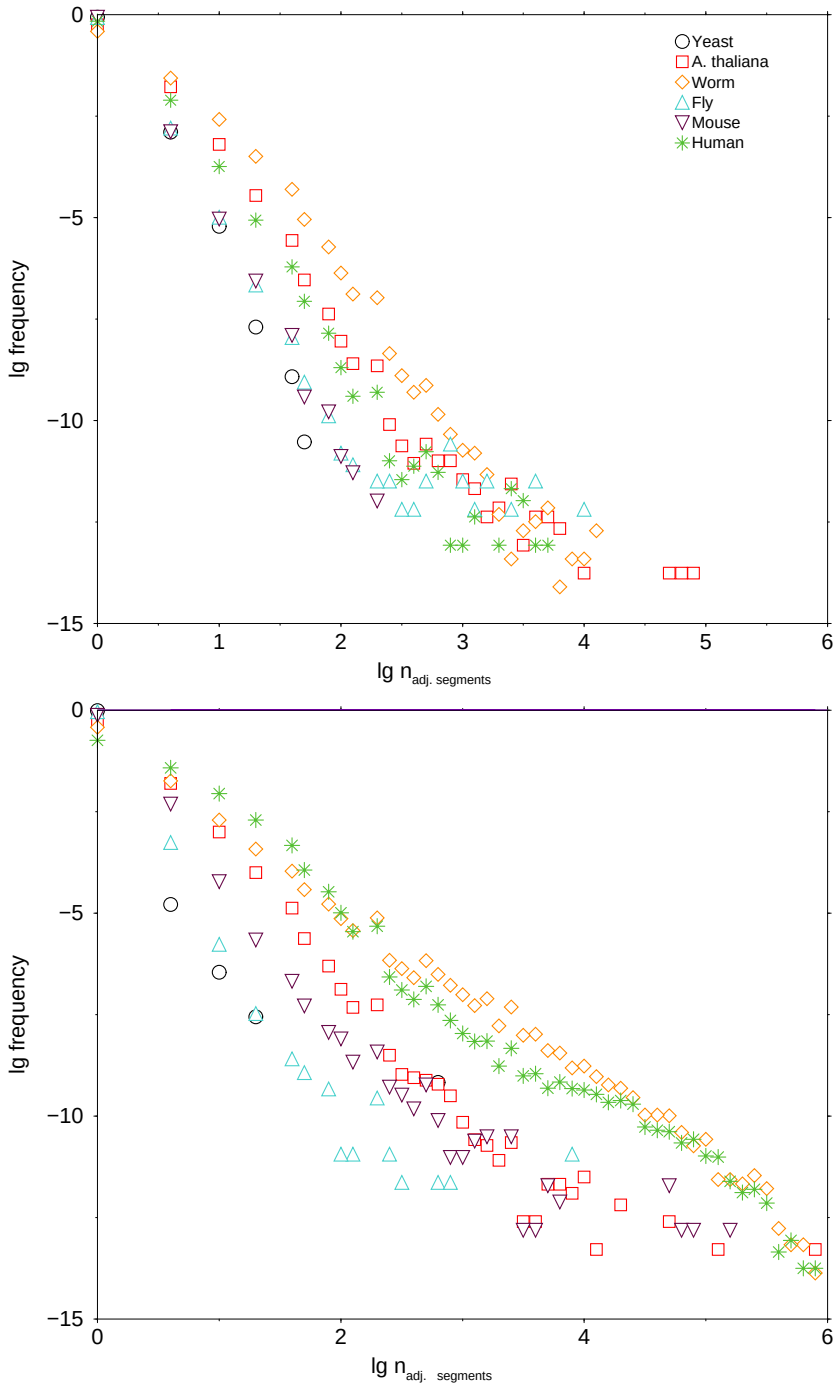


Figure 5.6: Connectivity distributions of exon (upper panel) and intron segments sized 11 nucleotides (lower panel) of Human, Fly, Worm, Mouse, Yeast and *Arabidopsis thaliana*. 10 % of each complete exon and intron sets were randomly picked. Exon and intron sets were retrieved from EID database.

5.4 DISCUSSION

5.4.1 Transition from Gaussian to power-law distribution

Interestingly enough, segmentations of sequences towards increasing size of segments and subsequent set up of graphs constitutes a shift in network topology which is indicated by a transition from a Gaussian distribution of connectivity to a real power law. This finding is accompanied by a sharp decline of the networks density. This observation holds for exon and intron sequences as well as for their randomized pendants.

The presence of power-laws in the connectivity distributions indicates the existence of scale-free networks which were initially found to be sparse but highly clustered networks. The emergence of scale-free networks is based on constant addition of new nodes which are preferentially attached to already well connected sites. Thus, a small subset of highly linked nodes determines the topology of the network. Accordingly, how does the presence of a transition from a Gaussian to a power-law distribution fit to this picture? The key lies in the densities of networks. Since it was observed that the Gaussian distribution of links corresponds to high network densities, the constant addition of new nodes appears to have a neglectable influence compared to preferential linkage of already existing nodes to present nodes. Obviously, this emphasize of preferential attachment with almost no growth depicts a limiting case of scale-free topology (Barabási et al. 1999).

Surprisingly, segmentations of all sequence samples proceed to power-law distributions although natural and randomized distributions of small segment sizes differ considerably. Presumably, this behavior is the consequence of the sample size of sequences. The number of possible combinations of nucleotides increases with 4^N and their pairwise combinations with 4^{2N} if N is the segment size. Thus, the capacity to find all combinations in a finite sample of sequences falls off rapidly with increasing N resulting in the random emergence of highly connected segments in the case of randomized sequences. In contrast, natural sequences, provide evolutionary significant segments which thus were copied and combined

repeatedly.

One might argue that investigations carried out on sets of randomly picked natural sequences might be an oversimplification. However, scale-free topology suggest networks which are actually independent of their network size. So, characteristics of networks will not change with increasing amount of information.

5.4.2 Divergence of network topologies

Obviously, relative entropies of networks set up by natural and randomized sequences reveals considerable differences over a broad range of segments. In the latter section, it was found that sufficiently small and large segments generate very dense and sparse graphs, respectively. Thus, these degrees of network density cause these networks to adopt very similar topologies. However, the broad range between segment sizes 4 and 12 shows considerable divergence between natural and random networks which indicates concise differences in the underlying network shapes. Coincidentally, these particular set of segment sizes covers roughly the section of transition. Obviously, this area seems to be the appropriate section to investigate the differences of sequence compositions between natural and random sequences as well as between exon and intron sequences. Regarding Figure 5.4, it is obvious that the sequence composition of exon sequences observed regardless of the organism is far more regular than the respective one of introns. Thus, it might be reasonable to argue that introns underwent different compositional procedures than exons.

5.4.3 Classification of segments

Segments of different length raise the hopes to find some of them exclusively in exon or intron sequences. Results clearly indicate that increasing segment size enhances the probability that a distinct segment belongs to the set of exons. The classification rule also takes the relative sizes of the exon and intron sets into account. It is well known that the size of these sets can differ tremendously between organisms. However, considerable size effects diminish fast with increasing

size of segments. Essentially, these findings also hold for randomized sequences. Since the random generation procedure ensures a broader variety of sequential composition, the convergence of classification runs even faster.

A good classification that a certain segment occurs either in exons or introns, respectively, can be achieved with reasonable small segment sizes. So, these small segments thus classified might be used as probe for the detection of relevant coding and noncoding sequences. Furthermore, segments could also be set up as combinations which might find relevant coding and noncoding sequences more efficiently and properly.

5.4.4 Evolutionary aspects

Since scale-free networks grow continously by constant addition of new nodes which are preferentially attached, these networks remain sparse but preserve a high degree of local clustering. This latter observation immediately suggests a core of nodes which turn out to be essential for the evolution of the network. Metabolic networks and domain networks identified such centers which proved to be the starting points for the evolution of metaboloms and proteomes (Jeong et al. 2000; Wagner and Fell 2001; Wuchty 2001). Similar findings also hold for those networks which segments are sufficiently long in order to gain scale-free topology. However, in this analysis segments do not fulfill the requirements of biological entities as is the case in the mentioned examples. Although the frame applied is clearly generic and does not mediate immediate biological meaning, the latter statements can still be made. Obviously, cores in networks of segments are spanned by repeatedly copied and combined segments. The idea that a certain biological entity such as a gene or domain is copied, subsequently modified and recombined is a very familiar pattern of genome and proteome evolution. These considerations about segment cores imply the perception that increasing sparseness of networks is accompanied by a decreasing degree of local clustering towards larger segments. Thus, cores diminish as the size of segments grows. Hence, significant cores presumably might be obtained with segment sizes which already were found to set up highly divergent network topologies (Figure 5.4).

Revisiting the idea of using segments as probes, core segments which also prove to be highly associated to exon sequences as well as their combinations might be the best choice for investigating otherwise unknown genomic sequences in terms of detecting genes.

Connectivity distributions of networks which were obtained from exon and intron sequences of some eukaryotes also have to be considered from the perspective of evolution. Figure 5.6 displays connectivity distribution of networks set up by 11-mers of exons and introns which were obtained from eukaryotic organisms. This particular length was chosen for clarity reasons. Especially, this particular segment size emphasizes strongly connectivity distributions which differ in their slopes. A similar result was obtained for connectivity distributions of domain networks. Due to the organisms complexity, different slopes were obtained (Wuchty 2001). Essentially, the repeatedly copying segments coincides with the increase of connectivity of the respective segment. Similarly to the results of proteomes, different slopes indicate the combinatorial requirements in order to set up sets of exons which are capable to maintain all cellular aspects of the underlying organism sufficiently. In order to avoid dramatical expansions of genomes, exon shuffling enabled the combinatorial emergence of different genes. Although segments are a rough abstraction of genetic entities evolution acts on, highly linked segments could be treated as evolutionary hubs which help to organize the genomic space by occasionally linking them to numerous other related segments. However, this observations also hold for introns albeit their distributions are far more fuzzy than the exons ones. Furthermore, distributions of introns decline slowlier. Thus, these observations might indicate the possibility that introns as well as exons undergo these evolutionary constrains although introns are currently not assumed to be that essential for the survivability of the cell. However, the result that introns of higher eukaryotes generate shallow connectivity distributions of segments might be a combinatorial consequence of their tremendous amount of intron sequences.

5.4.5 Revisiting former investigations

The results so far clarify some reports of recent time. If Sandberg et al. report a better performance while classifying sequences by segments of increasing size (Sandberg et al. 2001) this event coincides with the transition to a scale-free regime in an analogously generated network. As already shown, this observation accompanies improving quality of segments classification. Thus, the naive classifier which is based on the predictive power of Bayesian probabilities allows better prediction if centers of high probability i.e. highly connected segments, start to evolve with increasing length of segments.

Recent works emphasized the occurrence of di-, tri- and tetranucleotides which would set up regular graphs or at least networks providing a Gaussian distribution of links. It would be far presumptuous that these methods detect random noise albeit the scope was precisely the discovery of sections which clearly exaggerate noise and provide a characteristic genomic signature of the underlying organism. The results so far suggest that networks set up by segments of latter lengths do not exhibit scale-free properties. Since it was already found that certain ranges of segment sizes causes network topologies to diverge significantly and improves the quality of classification, it would be reasonable to apply the methods addressed to these values.

Investigating the abundance of di-, tri- and tetranucleotides, it was found that intergenomic differences are higher than intragenomic ones. It might be expectable that a comparison of genomes applying the approaches and methods which were introduced in this paper would yield new interesting results.

CHAPTER 6

Small Worlds in RNA

6.1 Introduction

Structures of RNA molecules can be discussed at an empirically well established level of resolution known as secondary structure which rather refers to a topology of binary contacts that arise from specific base pairings (Watson-Crick and GU, see Figure 6.1) than a geometry cast in terms of coordinates and distances. The driving force behind secondary structure formation is the stacking of contiguous base pairs. However, any formation of an energetically favorable double-stranded region implies the simultaneous formation of an energetically unfavorable loop. This 'frustrated' energetics lead to vast combinatorics of helix and loop arrangements which span the structural repertoire of an individual RNA sequence.
A secondary structure can be conveniently discretized as a graph which represents a pattern of base pair contacts (Figure 6.1). This yields a formally well-defined combinatorial object that can be subject to mathematical treatment. Of particular interest are secondary structures which satisfy some extremal condition, such as having largest number of admissible base pairs or minimal free energy. Structures of these kinds can be computed by dynamic programming (Nussinov et al. 1978; Nussinov and Jacobson 1980; Waterman and Smith 1978; Zuker and

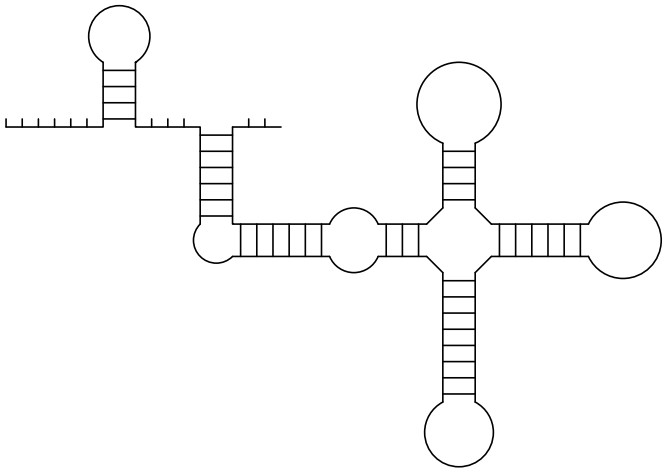


Figure 6.1: A RNA secondary structure graph. Unpaired positions not enclosed by base pairs, such as free ends or links between independent structure modules, are called 'external'.

Stiegler 1981; Zuker and Sankoff 1984; Hofacker et al. 1994; Zuker 2000). RNA thermodynamic folding algorithms have been extended to compute all suboptimal conformations (Wuchty et al. 1999; Zuker 1989).

The kinetics of RNA folding are controlled by the structure of the underlying free-energy landscape. These were recently investigated using folding algorithms capable of tracing the kinetic folding trajectories of RNA (Flamm et al. 2000; Isambert and Siggia 2000). Theoretical calculations predict some cases which indicate low barrier heights in free-energy landscapes of RNA (Flamm et al. 2002). Thus, the diffusion of a RNA's conformation on its energy landscape may be determined mainly by the structure and connectivity of its conformational space. Efforts have been done to model and investigate the structure and properties of the conformational space (Flamm et al. 1999).

All suboptimal RNA secondary structures within a certain energy range above the minimum free energy were computed in order to investigate the statistical properties of tRNA structures. It was found that base modification considerably sharpens the definition of the ground state structure by constraining energetically adjacent structures to be adjacent to the ground state (Wuchty et al. 1999).

Most recently, the conformational space of a simple lattice polymer chain was mapped to a network (Scala et al. 2001). Thus, conformations are connected if they switch by a single step out of a predefined move set of elementary conformational changes. The geometric properties of the network were found to be similar to those of small-world networks.

Since it was found that tRNA sequences and related sets of suboptimal structures have uniform properties, I consider the conformational spaces of a typical RNA sequence, *E.coli* tRNA^{phe}. In order to study the properties of tRNA^{phe}, the conformational spaces of the naturally and randomly modified and unmodified sequences will be mapped onto networks. Subsequently, the topology of these networks will be investigated and results discussed.

6.2 Material and methods

6.2.1 Secondary structures

A RNA sequence is denoted by a string $\mathsf{I} = (x_1, x_2, ..., x_n)$ of n positions over the familiar nucleotide alphabet, $x_i \in \mathcal{A} = \{\mathsf{A}, \mathsf{U}, \mathsf{G}, \mathsf{C}\}$. The bases x_1 and x_n are the nucleotides at the 5' and 3' ends, respectively. The usual formalization (Waterman 1995; Zuker and Sankoff 1984) views a secondary structure $\mathcal{S}$ as a graph (Figure 6.1) whose nodes represent nucleotides at positions $i = 1, \ldots, n$ of an RNA sequence of length n. The set of edges connecting the nodes consists of two disjoint subsets. One is common to all secondary structure graphs, while the other is specific to each sequence. The common set represents the covalent backbone connecting node i with node $i+1$, $\forall i = 1, \ldots, n-1$. The sequence specific part consists of a set Π of edges $i \cdot j$, $\Pi = \{\, i \cdot j \mid i \neq j \text{ and } j \neq i+1 \,\}$, representing admissible hydrogen bonds between the bases at positions i and j, such that (i) every edge in Π connects a node to at most one other node, and (ii) the pseudoknot constraint is met. The latter states that if both $i \cdot j$ and $k \cdot l$ are in Π, then $i < k < j$ implies that $i < l < j$. Failure to meet this constraint results in interactions which are considered to be tertiary contacts (pseudoknots). A sequence S is called compatible with a secondary structure $\mathcal{S}$, whenever positions that pair in the spec-

ification of $\mathcal{S}(i \cdot j \in \Pi(\mathcal{S}))$ are occupied by nucleotides which can actually pair with each other: $i \cdot j \rightarrow [x_i\,, x_j] \in \mathcal{B} = \{\mathsf{AU}, \mathsf{UA}, \mathsf{GC}, \mathsf{CG}, \mathsf{GU}, \mathsf{UG}\}, \forall i \cdot j \in \Pi(\mathcal{S})$. In other words, the set of admissible base pairs which we shall consider consists of the Watson-Crick pairs {AU,UA,GC,CG} and {GU,UG}. A sequence I specifies a set of structures S with which it is compatible, $\mathsf{S}(\mathsf{I}) = \{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_m\} \cup \{\mathsf{0}\}$, where $\mathcal{S}_0$ is the minimum free energy structure (mfe) and $\mathcal{S}_1, \ldots, \mathcal{S}_m$, are suboptimal conformations ordered with respect to their energy. 0 denotes the open chain conformation.

6.2.2 RNA folding algorithms

Considering all possible secondary structures one RNA sequence of length n can adopt, one would roughly estimate the total number of structures to $S_n \approx n^{-3/2} \times 1.85^n$ (Schuster et al. 1994). Thus, it is computationally expensive to calculate all secondary structures and henceforth to set up the respective conformational space. So, the set of conformations was restricted to suboptimal structures within a certain energy range above the minimum free energy structure. These structures were easily computed with the program `RNAsubopt` (Wuchty et al. 1999) which is part of the `Vienna RNA package` (http://www.tbi.univie.ac.at/ ivo/RNA/) (Hofacker et al. 1994). Essentially, `RNAsubopt` extends the standard RNA folding algorithm which emphasizes dynamic programming by an extended backtracking procedure. This admits the alternative arrangement of secondary structure elements in conformations which are within a certain energy range above the energy of the energetically most stable structure.

In order to obtain and investigate structures which are local energy minima and saddle points which connect minima by a downhill walk starting from them, the program `barriers` was used (http://www.tbi.univie.ac.at/ ivo/RNA/Barriers/) (Flamm et al. 2002). Taking the output of `RNAsubopt`, `barriers` starts to scan the vicinity of the minimum free energy structure in order to detect adjacent structures which might either be transient structures on the way to other optima, local minima or saddle points connecting them. This procedure is repeated for

the whole set of suboptimal structures. Energy barriers are the energy difference between a local minima to its saddle points.

6.2.3 Conformational space

The set S forms the conformational space $G_C(V_C, E_C)$ which considers its structures as the set of vertices V_C. The set of undirected edges E_C represents elementary moves between conformations which are restricted to the formation, removal and shift or flip move of base pairs. Figure 6.2 gives a schematic overview of these conformational changes. These can be considered from the perspective of a base pair based metric. Removal and formation of a base pair would cause $d = 1$ as the base pair distance between the respective merging conformations since one base pair is immediately affected. Analogously, a shift or flip move of base pairs would result in base pair distance $d = 2$, since two base pairs have been changed in the underlying conformation.

In this graph, the degree k_i of a vertex i is the number of other vertices to which it is linked. In other words, k_i represents the number of adjacent structures which are reachable with one elementary move from the given one.

The mean path length L from a vertex to any other vertex of the graph is defined as the average of the path lengths to all other vertices.

Another important quantity is the clustering coefficient C_i of a vertex i. It measures the fraction of the vertices connected to v which are also connected to each other. In extension, the clustering coefficient C of the graph is defined as the average of C_i over all i.

6.2.4 Graph tools

Graph analysis tools were written in C++ using the LEDA library of data types (Mehlhorn and Naeher 1999). PAJEK (the Slovene word for spider), a program for large network analysis and visualization, was used for the illustra-

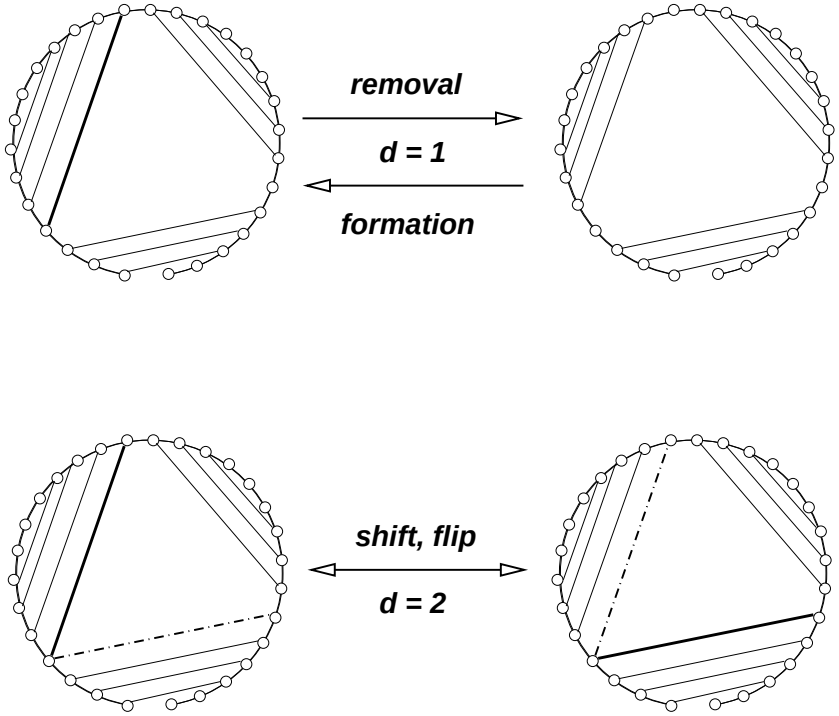


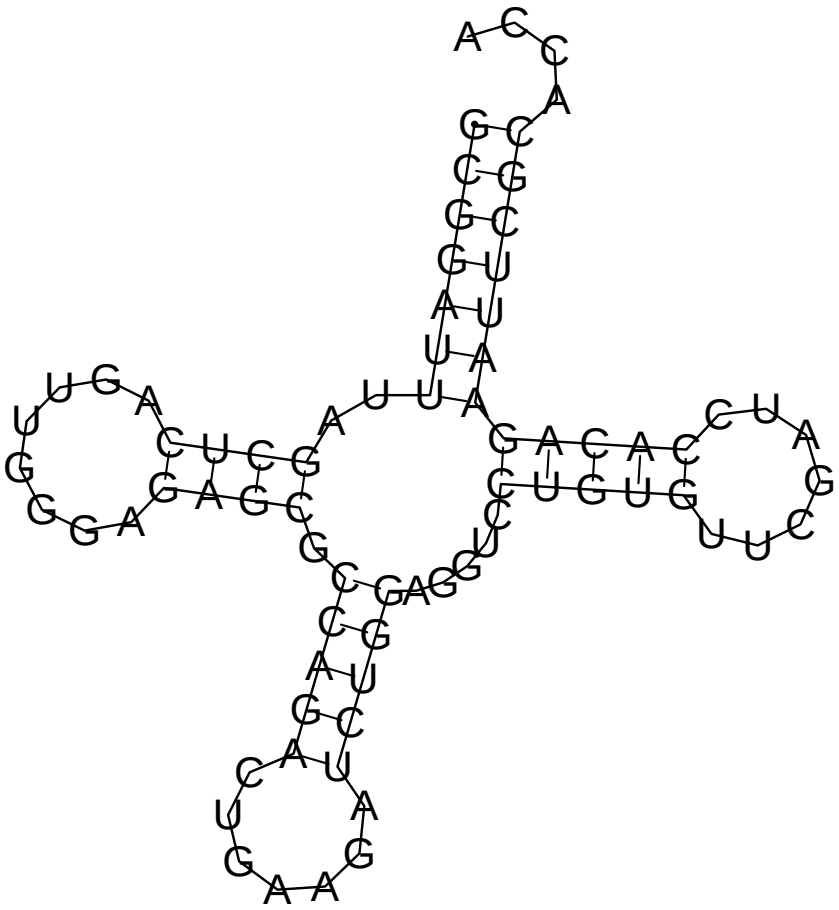
Figure 6.2: Elementary moves in the conformational space of RNA. Secondary structures are shown in circle representation. Base pairs which are going to change are indicated bold. Base pairs after a move are shown dot-dashed. Removal and formation of a base pair cause $d = 1$ as the base pair distance between the conformations since one base pair is immediately affected. Analogously, a shift or flip move of base pairs results in base pair distance $d = 2$, since two base pairs have changed the underlying conformations.

tions of the graphs (Batagelj and Mrvar 1998) (available at http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

6.2.5 tRNA sequences

It was already found that tRNA sequences constitute similar statistical properties (Wuchty et al. 1999). Hence, the *E.coli* tRNA^{phe} sequence (EMBL acc.no. RF6280) was exemplarily chosen as a typical protagonist of tRNAs from the compilation of Sprinzl et al. (http://www.uni-bayreuth.de/departments/biochemie/sprinzl/trna/) (Sprinzl et al. 1998). Bases are translated as suggested by Higgs (Higgs 1995). Some of these modifications still prevent the respective bases from pairing. The same number of such modifications was randomly distributed over the sequence in order to get the randomly modified sequence. Figure 6.3 gives a schematic

impression.



tRNA^{phe}:

naturally modified:

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

unmodified:

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

randomly modified:

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

Figure 6.3: Secondary structure of tRNA^{phe} (RF6280) and its sequences used throughout the analysis. The structure was obtained with the `RNAfold` algorithm. Modified bases are indicated red.

Sets of suboptimal structures within an energy range of $11\,kT$, where $T = 310.15K$, above the respective minimum free energies of all sequences were computed and conformational spaces thus obtained.

6.3 Results

The conformational spaces of the three tRNA sequences are sparse with small average degrees compared to the maximal possible degree $k = n - 1$ where n is the number of vertices (Table 1). Although just fractions - albeit the most important ones since they provide the global and the lowest local energy minima - of each conformational space is considered, the number of structures differs strongly (Table 6.1). Obviously, this is a consequence of modifications which restricts the number of structures and conformational transitions. The comparison of properties of the conformational graphs and respective random graphs of equal size reveals interesting results. Joint aspects will be discussed first.

	natural	none	random
n_v	1163	5853	2531
$\langle k_v \rangle$	9,77	13,15	10,54
$n_{conn.comp.}$	21	167	69

Table 6.1: Basic data of the conformational spaces of the naturally modified, unmodified and randomly modified *E.coli* tRNA^{phe} sequence.

	L	L_{random}	C	C_{random}
natural	5,72	2.70	0.5564	0.0161
none	6.85	2.94	0.1660	0.0016
random	5.89	2.86	0.2504	0.0044

Table 6.2: Characteristic path lengths, C, and mean path length, L, of the naturally modified, unmodified and randomly modified *E.coli* tRNA^{phe} sequence.

Mean path lengths L of all conformational spaces considered show that values of rough equal size and always exceed the respective numbers of equal sized random networks slightly (Table 6.2). Furthermore, Figure 6.4 shows a correlation

between the number of nodes, N, and the logarithm of the mean path lengths, $L \sim \log N$, of the underlying conformational networks which is typical for random graphs (Bollobás 1998) as well as for small-world networks (Barabási et al. 1999). Regarding Figure 6.4, it is obvious that this correlation only holds for

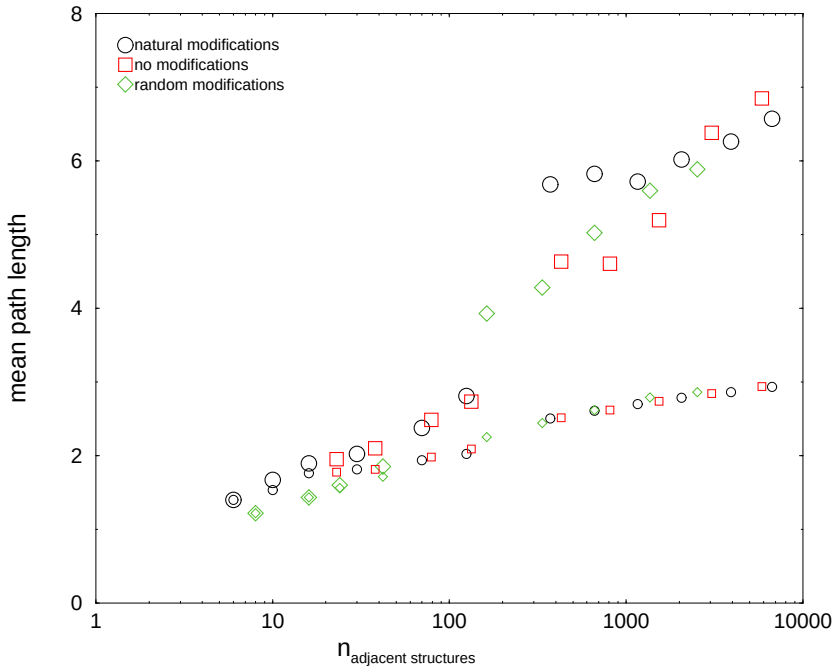


Figure 6.4: Total number of structures vs. mean path length of the underlying conformational space. Modified, unmodified and randomly modified sequences of *E.coli* $\mathrm{tRNA^{phe}}$ were considered. Large symbols refer to these data points. Suboptimal structures lie within $11kT$ above the respective minimum free energy. Analogously, the respective numbers of random graphs of equal size were plotted which refer to the small symbols.

a sufficient large number of nodes and is roughly independent of the modifications superimposed onto the tRNA sequence. Also the respective numbers of the random graphs of equal size were plotted which show the logarithmic correlation too. However, the slope of the curves is far lower than the ones of the natural occurring graphs. Proceeding with joint features of these graphs, Figure 6.5 shows connectivity distributions of the conformational spaces considered. Regardless of the tRNA modifications, the exponentially decaying connectivities remind to Poisson distributions which are typical for random graph topology as well as for small-world networks. Differences between conformational and random graphs are intriguing with respect to mean clustering coefficients, C. Table 6.2

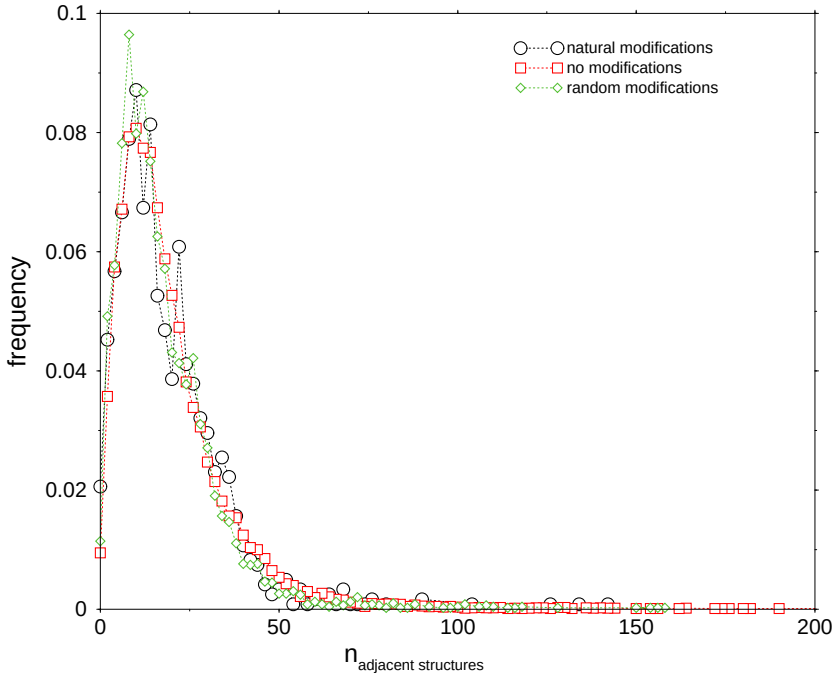


Figure 6.5: Connectivity distribution of the conformational spaces. Modified, unmodified and randomly modified sequences of *E.coli* $tRNA^{phe}$ were considered. Suboptimal structures lie within $11kT$ above the respective minimum free energy. Connectivity numbers were binned and frequencies thus obtained.

shows that the numbers regarding the conformational spaces greatly exceed the respective ones of the random graphs. This portrays the existence of small-world topology in conformational spaces of RNA. Furthermore, it is important to note that this finding is apparently independent of the degree of modification.

Considering properties which are characteristic for small-world networks, the mean path length, L, and the mean clustering coefficient, C, should be compared. Figure 6.6 shows histograms of these properties regarding the conformational spaces under consideration. Table 6.2 suggested the conformational space of the modified sequence to be distinctively more clustered than the other ones. However, it should be noted that in comparison to the unmodified sequence the randomly modified generates a more clustered conformational space. These observations will go to more detail in the histograms of Figure 6.6. Obviously, the distinct natural modifications of the tRNA shift clustering coefficients, C_v, to

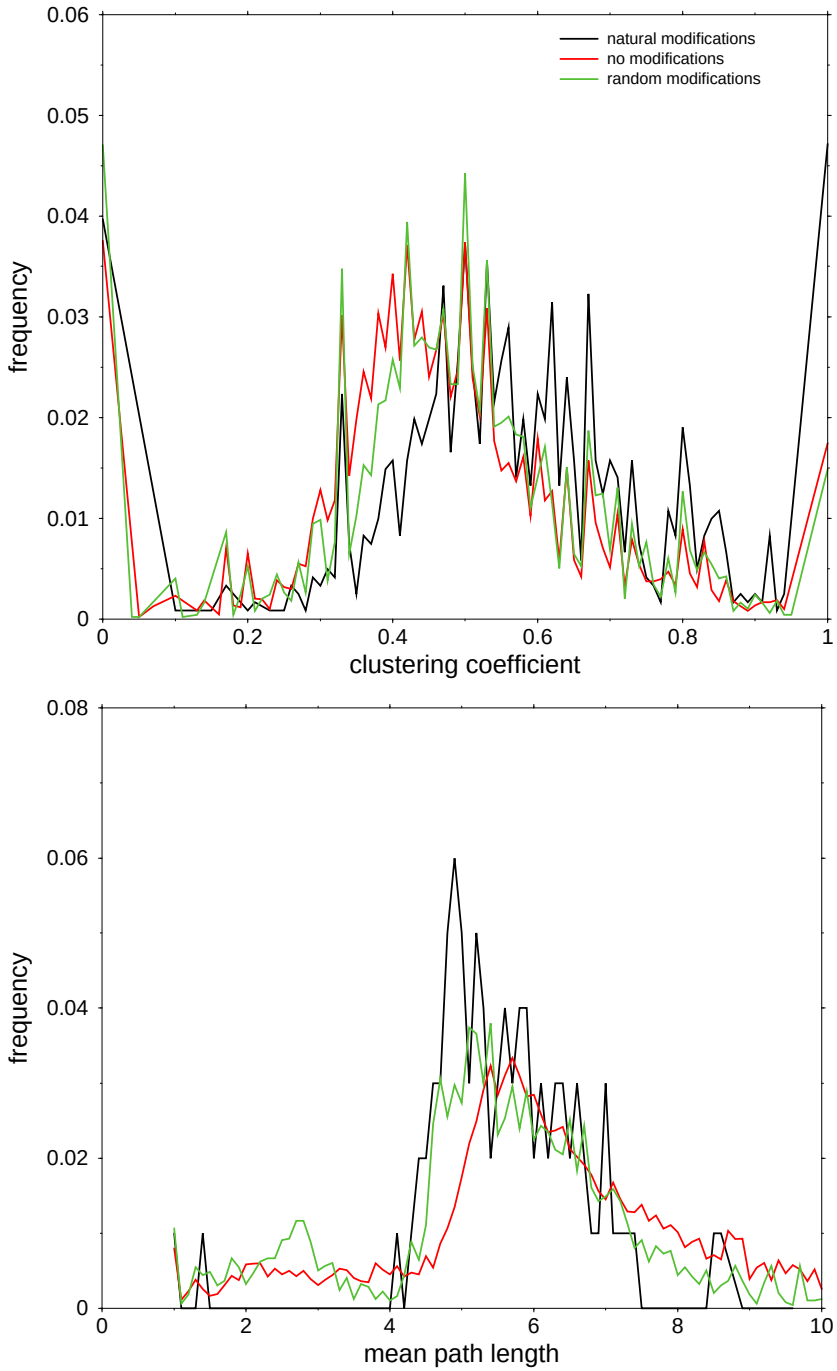


Figure 6.6: Histograms of mean clustering coefficient, C, and mean path lengths, L. Conformational spaces of modified, unmodified and randomly modified sequences of *E.coli* tRNA^{phe} were considered.

higher values. In principle, this observation also holds for the numbers of the randomly modified sequence. On the other hand, Table 6.2 suggests mean path lengths, L_v, of the modified sequences to lower numbers. This trend is confirmed in Figure 6.6 which indicates that natural modifications drive the distribution of L to smaller values. Furthermore, the latter distribution proves to be narrower than the respective ones of conformational spaces which were generated by randomly and unmodified sequences.

In a Boltzmann weighted ensemble, the probability of the ith suboptimal structure to occur is defined as

$$p_i = \frac{e^{-E_i/kT}}{\sum_i e^{-E_i/kT}}, \tag{6.1}$$

where the sum over all suboptimal structures defines the partition function.

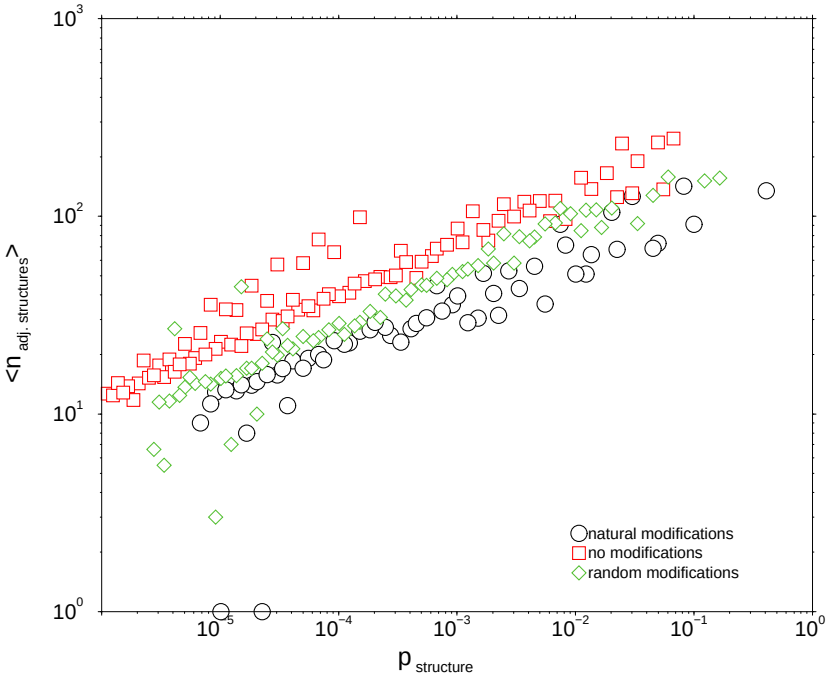


Figure 6.7: Probability of a certain structure against arithmetic mean number of adjacent structures. Conformational spaces of modified, unmodified and randomly modified sequences of *E.coli* tRNA^{phe} were considered.

Figure 6.7 and 6.8 show some correlations which are related to this ensemble probability of structures. Interestingly, regardless of the degree of modification all conformational spaces show a positive power-law dependency of the probability of structures to the arithmetic mean number of transitions to other structures.

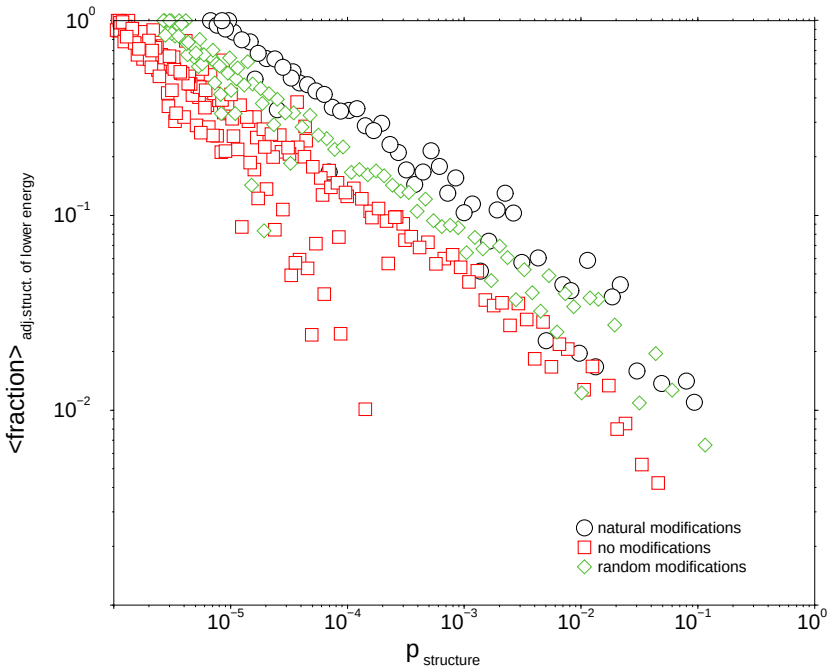


Figure 6.8: Probability of a certain structure against arithmetic mean fraction of links which point to structures of lower energy. Conformational spaces of modified, unmodified and randomly modified sequences of *E.coli* tRNA^{phe} were considered.

In other words, the more probable (i.e. stable in thermodynamic terms) the structure in the ensemble, the higher the probability to find it frequently in the vicinity of other structures. In contrast, the more frequent one structure occurs in the ensemble, the lower the probability to find a structure of lower energy in its vicinity which is indicated by its mean fractions. Since modifications prevent the bases affected from binding, a considerable amount of conformations can not occur which shifts distributions of modified sequences to higher ensemble probabilities. Interestingly, power-law coefficients prove to be approximately the same for all distributions. So, modifications prevent conformations from folding but leave the structure of the underlying conformational spaces essentially unchanged. In order to construct a more comprehensive image of the conformational spaces, Figure 6.9 takes a look on the impact of modifications to barrier heights which separates local minima energetically. Interestingly, a partly exponential behavior can be found in these frequency distributions. Nevertheless, Figure 6.9 suggests that the distribution of barrier heights is not influenced saliently by tRNA modifications

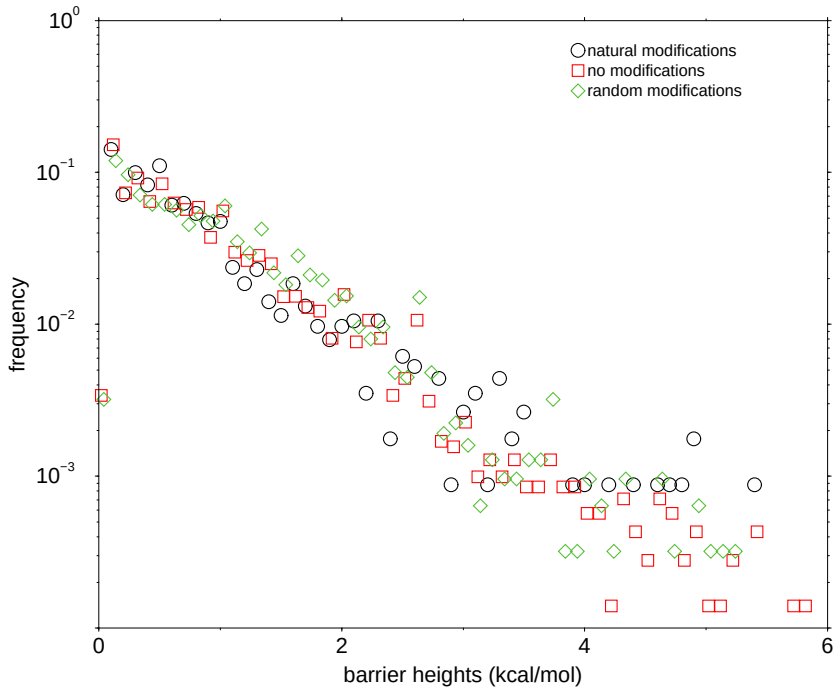


Figure 6.9: Frequency distributions of barrier heights which occur in conformational spaces of modified, unmodified and randomly modified sequences of *E.coli* tRNA^{phe}.

since all distribution tend to coincide in ranges of high frequency.

6.4 Discussion

6.4.1 Data of conformational spaces

It was already mentioned that the total number of structures estimated was tremendous. Thus, I only considered cutouts of the respective conformational spaces. Table 6.1 shows that these small cutouts comprise high numbers of connected components which are strongly subjected to modifications. One might argue that this perspective on conformational spaces might be too simple. However, the complete knowledge of the whole structure space is not necessary in order to comprehend its fundamental properties since it was explicitly shown that modifications which considerably reduces the number of structures do not change the underlying topology of the conformational space.

6.4.2 Aspects of small-worldedness

The topology of small-world networks uncovers nodes which prove to be more highly linked than the average nodes. In other works, these particular nodes were credited a special role. This particular set of nodes were identified as an evolutionary 'core' in metabolic networks (Fell and Wagner 2000; Wagner and Fell 2001). Similarly, some protein domains were found to serve as starting points of proteome evolution (Wuchty 2001). Obviously, highly linked structures can be credited a similar role. They prove to be energetically stable and frequently occurring in the thermodynamic ensemble which denote a local minima in a reasonable amount of cases. Since these conformations show high numbers of structures in their vicinity, they have to comprise a structural and energetical disposition which enable them to transit from one structure to the other. However, it should also be kept in mind that this observation is subject to the move set which essentially shapes the conformational space.

The most salient features of small-world topology in conformational spaces of RNA is the high degree of local clustering. Results indicate that natural modifications influence the degrees of local clustering far more than random or no modifications. The immediate result is a shift towards high numbers of the mean clustering coefficient, C (Figure 6.6). Thus, the natural modifications have a subtle streamlining effect on the shape of the conformational space.

The second feature of small-worldedness is the mean path length through a network, L. The influence of the topology on L is not as thorough as it is on the mean clustering coefficient. However, subtle differences between the degrees of modifications can still be detected. Coincidently, high degree of local clustering is accompanied by a streamlined mean path length distribution. Obviously, this finding is also subject to modifications.

To conclude, modifications leave the nature of small-world topology untouched albeit natural modifications have a reasonable enhancing and streamlining effect on the features of this topology.

6.4.3 Imagination of energy landscapes

Modifications of any kind do not alter the barrier heights between local minima. However, the influence on the landscape is a different one. Since modifications inhibit the folding of some structures, the underlying conformational space remains more sparse than conformational spaces of unmodified sequences. Henceforth, modifications also prevent partly energetically unfavorable barriers from emergence leading to energy minima which might be traps for the folding process. Thus, conformational spaces of unmodified RNA sequences show a sculptured landscape with considerably more possibilities to slow the folding process by getting trapped preferentially in an energetically unfavorable energy minimum (Figure 6.10). The probability that the folding process stops in a particular minimum in this smooth landscape depends essentially on the barrier heights. The kinetic folding process can thus be simulated stochastically (Flamm et al. 2000) and is beyond the scope of this work. Focusing on conformational spaces which were shaped with the aid of modifications, the topology is substantially modified. Figure 6.11 sums the results of this study. Since the folding opportunities and transitions between them are substantially limited, certain folding funnels emerge and lead more frequently to distinct kinetically favorable structures. Considering stochastic simulations of tRNA conformation spaces, they frequently prove to be the biological meaningful ones (Flamm et al. 2000). This observation coincides with a increased degree of connectivity and local clustering which enhances the relevance of folding funnels and focuses the mean number of steps through the conformational space.

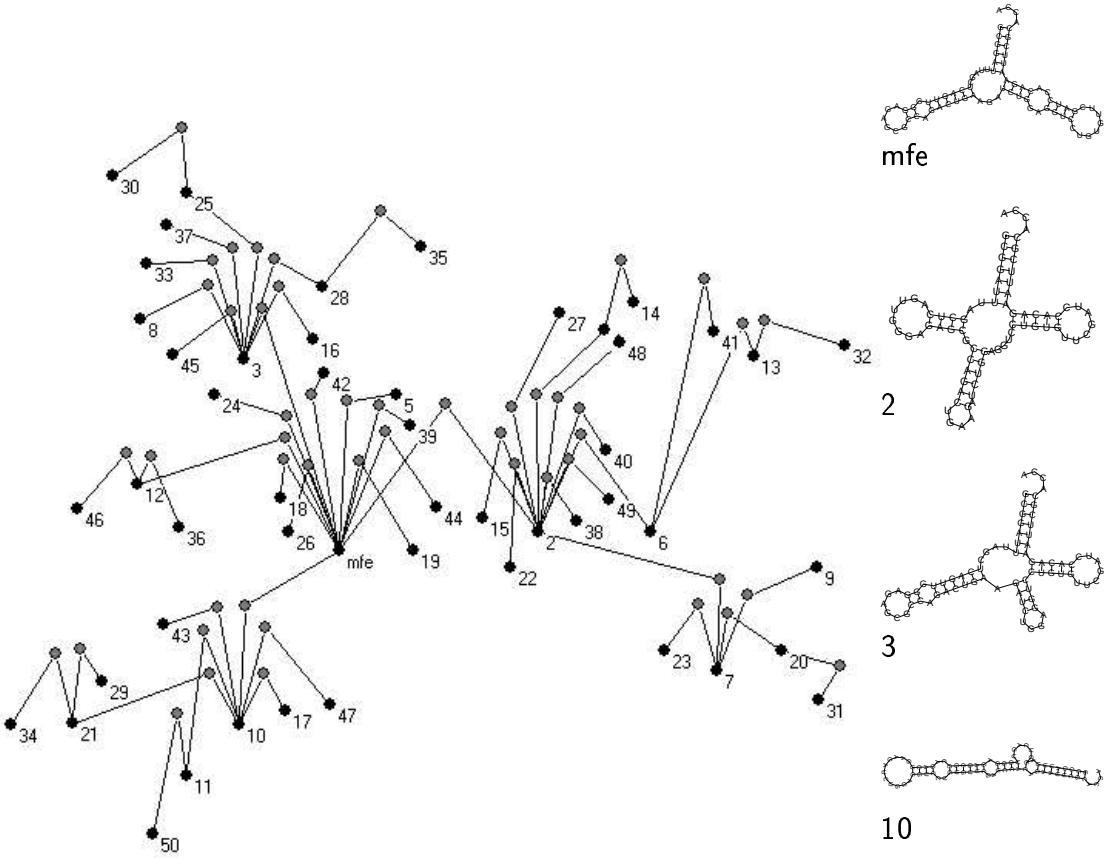


Figure 6.10: Merged landscape of the unmodified *E.coli* tRNA^{phe} sequence. Black dots characterize local minima, light grey ones denote saddle points. Secondary structures indicate the minimum free energy structure and structures of some important local minimas.

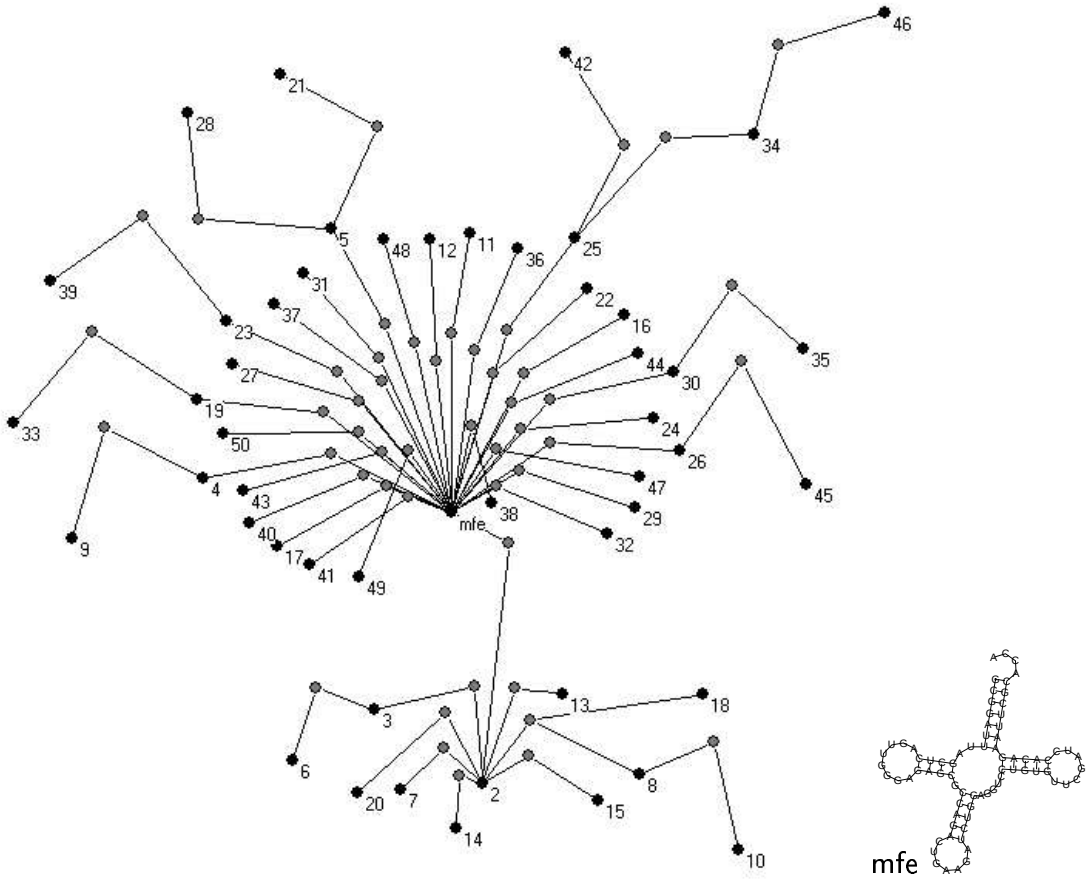


Figure 6.11: Merged landscape of the naturally modified *E.coli* $tRNA^{phe}$ sequence. Black dots characterize local minima, light grey ones denote saddle points. The secondary structure indicates the minimum free energy structure.

CHAPTER 7

Conclusions and outlook

The discovery of small-world and scale-free properties in various biological networks sheds a new light on the discussion about the significance of their topologies. Networks of protein domains, protein and domain interactions, genomic segments and RNA structure spaces turn out to be sparse but remain locally well clustered providing occasional long-range connections between these clusters.
The scale-free and small-world model explain the evolutionary emergence of these biological networks particularly well. In these models, evolutionary relevance is rewarded with an increasing degree of connectivity. Regarding scale-free topology, this is the immediate result of continuous addition of nodes and their subsequent preferential attachment. The set up of small-world networks emphasizes randomly rewiring of an initially regular graph as the most crucial point. In contrast to scale-free networks, the total number of nodes remains constant. Although the results are very encouraging, it has to be kept clearly in mind that both models are a considerable simplification of the real situation.

Domain distributions in proteomes reflect the underlying organisms complexity indicating an evolutionary trend to higher connectivity of domains towards multicellular eukaryotic organisms. Thus, the emergence of multicellularity re-

quires proteomes which feature complex cellular processes like signal transduction or cell-cell contacts. Expansion of particular domain families, domain accretion and extensive shuffling of domains leads to increasing combinatorial diversity which imitate the essential processes of the scale-free and small-world model. So, protein sets are provided which are sufficient to preserve cellular procedures without dramatically expanding the absolute size of the protein complement. Thus, highly connected domains display the emerging importance of the cellular processes mentioned and denote functional centers which contribute essentially to the large scale organization of the domain space.
In this study, the connectivity of nodes was taken as a measure of centrality. Since this value does not reflect the complexity of the network on the whole and the influence of a single node in particular, other measures of centrality might be applied. Such measures might be used as tools to dissect proteomes of organisms in order to identify the occurrence of key domains, their combinations and the emergence of their functions in evolution.

The comparison of protein, domain interaction and domain networks of *Saccharomyces cerevisiae* significantly shows that there is merely a weak correlation between lethality and transitivity of nodes as well as their degree of connectivity. Thus, the idea that lethal proteins and domains accumulated considerably more connections which was stated elsewhere recently does not hold. Furthermore, domain interactions do not prove to be the driving force of domain fusions.
In contrast to the interaction networks, domain networks have been found to exhibit considerable small-world properties which might have been the consequence of 'domain rewiring' processes during the proteomes evolution.
The set of protein interactions is determined by the set of domains the proteins provide. Since the amount of interaction data grows constantly and the quality of domain information still improves, both sources of information can be combined in order to provide a framework which enables the reliable prediction of interactions of otherwise unknown proteins.

A segmentation of sequences to equal lengths results in adjacent pieces which resemble a network structure. The consideration of a series of increasing segment

sizes constitutes a shift in network topology which is indicated by a transition from a Gaussian distribution of connectivity to a real power-law. This finding is obviously an immediate consequence of decreasing sparsity of networks. Interestingly, this observation holds for exon and introns as well as for randomly generated sequences albeit distributions of natural and random sequences differ significantly. Similarly to the domain networks, natural sequences provide evolutionary significant segments which thus were copied and combined repeatedly. So, the connectivity distributions reflect the biological complexity of the underlying organisms. In contrast to protein domains, segments are a rough treatment of biological entities.
Considering differences of network topologies which were set up by natural and random sequences, relative entropy as a measure of divergence was applied. Thus, a comparison proved that network topologies differ significantly with segment sizes ranging from 4 to 12. This areas seem to be the appropriate sections to investigate the differences of sequence compositions between natural and random sequences as well as between exon and intron sequences.
A reasonable classification that a certain segment occurs either in exons or introns, respectively, can be achieved with relatively small segment sizes which might be used as effective probes for the detection and investigation of exon and intron sequences.
The idea of segmenting sequences resembles the emergence of a language. However, segments of equal length are a very rough abstraction of words. In order to obtain a 'DNA language' that helps to get insights into e.g. sequential composition of exons and introns and protein structure, it is necessary to define new words of different lengths.

Conformational spaces of tRNA set up by sets of suboptimal structures and a move set defining transitions between them were discussed from the perspective of small-world topology. On the one hand, independent from any modification which inhibits distinct structures from folding, the conformational spaces adopt small-world topology which emphasizes local clusters of structures and a few transitions between them. On the other hand, modifications prove to enhance these typical small-world network properties. They influence the shape of the

energy landscape considerably showing a good correlation between mean numbers of adjacent structures and some degrees of connectivity in the underlying conformational space.

Summarizing, the study of the topologies of biochemical networks resulted in an insight into their evolution and function in a very effective way. The understanding, how networks of a certain topology emerge, gives us an increased knowledge about biological history. Moreover, functional aspects of information flow and robustness can be inferred.

APPENDIX A

References

Albert, R., and A. Barabási. 2000. Topology of evolving networks: local events and universality. Phys. Rev. Lett. **85**:5234.

Albert, R., and A. Barabási. 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. **74**:47.

Albert, R., H. Jeong, and A. Barabási. 1999. Diameter of the World Wide Web. Nature **401**:130–131.

Albert, R., H. Jeong, and A. Barabási. 2000. Error and attack tolerance of complex networks. Nature **406**:378–382.

Altschul, S., T. Madden, A. Schaeffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. **25** (17):3389 – 3402.

Amaral, L., A. Scala, M. Barthélémy, and H. Stanley. 2000. Classes of small-world networks. Proc. Natl. Acad. Sci. **97**:11149–11152.

Apic, G., J. Gough, and S. Teichmann. 2001. Domain Combinations in Archaeal, Eubacterial and Eukarytic Proteomes. J. Mol. Biol. **310** (2):311–325.

Apweiler, R., T. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I.

Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. Mulder, T. Oinn, M. Pagni, and F. Servant. 2001a. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucl. Acids Res. **29** (1):37–40.

Apweiler, R., M. Biswas, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E. Kriventseva, V. Mittard, N. Mulder, I. Phan, and E. Zdobnov. 2001b. Proteome Analysis Database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. Nucl. Acids Res. **29** (1):44–48.

Aravind, L., V. Dixit, and E. Koonin. 2001. Apoptotic Molecular Machinery: Vastly Increased Complexity in Vertebrates Revealed by Genome Comparisons. Science **291**:1279–1284.

Attwood, T., M. Croning, D. Flower, A. Lewis, J. Mabey, P. Scordis, J. Selley, and W. Wright. 2000. PRINT-S: the database formerly known as PRINTS. Nucl. Acids. Res. **28** (1):225–227.

Bairoch, A., and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucl. Acids Res. **28**:45–48.

Barabási, A., and R. Albert. 1999. Emergence of Scaling in Random Networks. Science **286**:509–512.

Barabási, A., R. Albert, and H. Jeong. 1999. Mean-field theory for scale-free random networks. Physica A **272**:173–187.

Barabási, A., R. Albert, and H. Jeong. 2000. Scale-free characteristics of random networks: the topology of the world-wide web. Physica A **281**:69–77.

Barabási, A., H. Jeong, R. Ravasz, Z. Neda, T. Vicsek, and A. Schubert. 2002. On the topology of the scientific collaboration networks. Physica A, in press.

Barrat, A., and M. Weigt. 2000. On the properties of small-world network models. Eur. Phys. J. B. **13**:547.

Barthélémy, M., and L. Amaral. 1999. Small -World Networks: Evidence for a Crossover Picture. Phys. Rev. Letters **82** (15):3180–3183.

Batagelj, V., and A. Mrvar. 1998. PAJEK - Program for Large Network Analysis. Connections **21** (2):47–57.

Bateman, A., E. Birney, R. Durbin, S. Eddy, K. Howe, and E. Sonnhammer. 2000. The Pfam Protein Families Database. Nucl. Acids Res. **28**:263–266.

Bianconi, G., and A.-L. Barabási. 2001. Competition and multiscaling in evolving networks. EuroPhys. Lett. **54**:436.

Blaisdell, B., A. Campbell, and S. Karlin. 1996. Similarities and dissimilarities of phage genomes. Proc. Natl. Acad. Sci. USA **93**:5854–5859.

Bollobás, B. 1998. Modern Graph Theory. Springer, New York.

Bornberg-Bauer, E. 1997. How are model protein structures distributed in sequence space? Biophysical J. **5** (73):2393–2403.

Burge, C., A. Campbell, and S. Karlin. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA **89**:1358–1362.

Corpet, F., F. Servant, J. Gouzy, and D. Kahn. 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucl. Acids. Res. **28** (1):267–269.

Costanzo, M., M. Crawford, J. Hirschmann, J. Kranz, P. Olsen, L. Robertson, M.S.Skrzypek, B. Braun, K. Hopkins, P. Kondu, C. Lengieza, J. Lew-Smith, M. Tillberg, and J. Garrels. 2001. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge Library, an integrated resource for protein information. Nucl. Acids Res. **29** (1):75–79.

de Menezes, M. A., C. Moukarzel, and T. Penna. 2000. First-order phase transition on small-world networks. Europhys. Lett. **50** (5).

Deschavanne, P., A. Giron, A. Vilain, G. Fagot, and B. Fertil. 1999. Genomic signature: Characerisation and classification of species assessed by chaos game reprentation of sequences. Mol. Biol. Evol. **16**:1391–1399.

Doolittle, R. 1995. The Multiplicity of Domains in Proteins. Ann. Rev. Biochem. **64**:287–314.

Dorit, R., and W. Gilbert. 1991. The limited universe of exons. Curr. Opin. Genet. Dev. **1** (4):464–469.

Dorogovtsev, S., and J. Mendes. 2000a. Exactly solvable analogy of small-world networks. Europhys. Lett. **50**:1.

Dorogovtsev, S., and J. Mendes. 2000b. Evolution of reference networks with aging. Phys. Rev. E **62**:1842.

Dorogovtsev, S., and J. Mendes. 2000c. Scaling behaviour of developing and decaying networks. Europhys. Lett. **53**:33.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis. Cambridge University Press, Cambridge, England.

Erdös, P., and A. Rényi. 1960. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. **5**:17–61.

Fell, D., and A. Wagner. 2000. The small world of metabolism. Nature Biotech. **189**:1121–1122.

Fickett, J., and C.-S. Tung. 1992. Assessment of protein coding measures. Nucl. Acids Res. **20** (4):6441–6450.

Flamm, C., W. Fontana, I. Hofacker, and P. Schuster. 2000. RNA Folding at Elementary Step Resolution. RNA **6**:325–338.

Flamm, C., I. Hofacker, and P. Stadler. 1999. RNA in silico: The Computational Biology of RNA Secondary Structures. Adv. Complex Syst. **2**:65–90.

Flamm, C., I. Hofacker, P. Stadler, and M. Wolfinger. 2002. Barrier Trees of Degenerate Landscape. Z.Phys.Chem. **216**:155–173.

Gentles, A., and S. Karlin. 2001. Genome-Scale Compositional Comparisons in Eukaryotes. Genome Res. **11**:540–546.

Gilbert, W., and M. Glynias. 1993. On the ancient nature of introns. Gene **135**:137–144.

Gleiss, P., P. Stadler, A. Wagner, and D. Fell. 2001. Relevant Cycles in Chemical Reaction Networks. Adv. Complex Systems **4**:207–226.

Grosse, I., H. Herzel, S. Buldyrev, and H. Stanley. 2000. Species independence of mutual information in coding and noncoding DNA. Phys. Rev. E **61** (5):5624–5629.

Guare, J. 1990. Six Degrees of Seperation: A Play. Vintage Books, New York.

Hazbun, T., and S. Fields. 2001. Networking proteins in yeast. Proc. Natl. Acad. Sci. **98**:4277–4278.

Herzel, H., E. Trifonov, O. Weiss, and I. Grosse. 1998. Interpreting Correaltions in Biosequences. Physica A **249**:449–459.

Herzel, H., O. Weiss, and E. Trifonov. 1999. 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics **15** (3):187–193.

Higgs, P. G. 1995. Thermodynamic properties of transfer RNA: A computational study. J .Chem. Soc. Faraday Trans. **91** (16):2531–2540.

Hofacker, I. L., W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. 1994. Fast folding and comparison of RNA secondary structures. Monatsh. Chem. **125** (2):167–188.

Hofmann, K., P.Bucher, L. Falquet, and A. Bairoch. 1999. The PROSITE database, its status in 1999. Nucl. Acids Res. **27**:215–219.

Holste, D., I. Grosse, S. Buldyrev, H. Stanley, and H. Herzel. 2000. Optimization of Coding Potentials Using Positional Dependence of Nucleotide Frequencies. J. Theor. Biol. **206**:525–537.

Huberman, B., and L. Adamic. 1999. Growh dynamics of the World -Wide Web. Nature **401**:131.

Huberman, B., P. Pirolli, J. Pitkow, and R. Lukose. 1998. Strong Regularities in World Wide Web Surfing. Science **280**:95–97.

i Cancho, R. F., and R. Solé. 2001. Optimization in complex networks. Santa Fe Institute Working paper 01-11-068.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Isambert, H., and E. Siggia. 2000. Modeling RNA Folding Paths With Pseudoknots: Application To Hepatitis Delta Virus Ribozyme. Proc. Natl. Acad. Sci. USA **97** (12):6515–6520.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Nat. Acad. Sci. **98** (8):4569–4574.

Ito, T., K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. 2000. Towards a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Nat. Acad. Sci. **97** (3):1143–1147.

Janin, J., and C. Chothia. 1985. Domains in proteins: definitions, location, and structural principles. Methods Enzym. **115**:420–430.

Jeong, H., S. Mason, A.-L. Barabási, and Z. Oltvai. 2001. Lethality and centrality in protein networks. Nature **411**:41–42.

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

KARLIN, S., and C. BURGE. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics **11** (7):283–290.

KARLIN, S., and I. LADUNGA. 1994. Comparisons of eukaryotic genomic sequences. Proc. Natl. Acad. Sci. USA **91**:12832–12836.

KARLIN, S., I. LADUNGA, and B. BLAISDELL. 1994. Heterogeneity of genomes: Measures and values. Proc. Natl. Acad. Sci. USA **91**:12837–12841.

KARLIN, S., and J. MRÁZEK. 1997. Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. USA **94**:10227–10232.

KARLIN, S., J. MRÁZEK, and A. CAMPBELL. 1997. Compositional Biases of Bacterial Genomes and Evolutionary Implications. J. Bacteriol. **179** (12):3899–3913.

KLEINBERG, J., and S. LAWRENCE. 2001. The Structure of the Web. Science **294**:1849–1850.

KRAPIVSKY, P., S. REDNER, and F. LEYVRAZ. 2000. Connectivity of Growing Random Networks,. Phys. Rev. Lett. **85**:4629–463.

KRIVENTSEVA, E., W. FLEISCHMANN, E. ZDOBNOV, and R. APWEILER. 2001. CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. Nucl. Acids Res. **29** (1):33–36.

LAWRENCE, S., and C. GILES. 1998. Searching the World Wide Web. Science **280**:98–100.

LI, W.-H., Z. GU, H. WANG, and A. NEKRUTENKO. 2001. Evolutionary analyses of the human genome. Nature **409**:847–849.

LILJEROS, F., C. EDLING, L. AMARAL, and Y. ABERG. 2001. The web of human sexual contacts. Nature **411**:907–908.

MANNING, C., and H. SCHÜTZE. 1999. Foundations of Statistical Language Processing. The MIT Press, Cambridge, USA.

MANTEGNA, R., S. BULDYREV, A. GOLDBERGER, S. HAVLIN, C.-K. PENG, M. SIMONS, and H. STANLEY. 1994. Linguistic Features of Noncoding DNA Sequences. Phys. Rev. Lett. **73** (23):3169–3172.

MANTEGNA, R., S. BULDYREV, A. GOLDBERGER, S. HAVLIN, C.-K. PENG, M. SIMONS, and H. STANLEY. 1995. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. Phys. Rev. E **52** (3):2939–2950.

Marcotte, E., M. Pellegrini, H.-L. Ng, D. Rice, T. Yeates, and D. Eisenberg. 1999. Detecting Protein Function and Protein-Protein Interactions from Genom Sequences. Science **285**:751–753.

Martindale, C., and A. Konopka. 1996. Oligonucleotide frequencies in DNA follow a Yule distribution. Computers Chem. **20** (1):35–38.

Mehlhorn, K., and S. Naeher. 1999. The LEDA Platform of Combinatorial Computing. Cambridge University Press, Cambridge.

Mewes, H., D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schüller, S. Stocker, and B. Weil. 2000. MIPS: a database for genomes and protein sequences. Nucl. Acids Res. **28** (1):37–40.

Milgram, S. 1967. The Small-World Problem. Psychology Today **2**:60–67.

Miller, G., and E. Newman. 1958. Tests of a statistical explanation of the rank-frequency relation for words in written English. Amer. J. of Psychol. **71**:209–218.

Nakashima, H., M. Ota, K. Nishikawa, and T. Ooi. 1998. Genes from nine genomes are separated into organisms in the dinucleotide composition space. DNA Res. **5**:251–259.

Newman, M. 2001. The structure of scientific collaboration networks. Proc. Nat. Acad. Sci. **98**:404–409.

Newman, M., S. Strogatz, and D. Watts. 2001. Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E **64**:026118.

Newman, M., and D. Watts. 1999. Renormalization group analysis of the small-world network model. Phys. Lett. A **23**:7332–7342.

Nussinov, R., and A. B. Jacobson. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc. Natl. Acad. Sci. USA **77** (11):6309–6313.

Nussinov, R., G. Piecznik, J. R. Griggs, and D. J. Kleitman. 1978. Algorithms for loop matching. SIAM J. Appl. Math. **35** (1):68–82.

Park, J., M. Lappe, and S. Teichmann. 2001. Mapping Protein Family Interactions: Intramolecular and Intermolecular Protein Family Interaction Repertoires in the PDB and Yeast. J. Mol. Biol. **307**:919–938.

Peng, C.-K., S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. Stanley. 1992. Long-range correlations in nucleotide sequences. Nature **356**:168–170.

Rain, J.-C., L. Selig, H. DeReuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schächter, Y. Chemama, A. Labigne, and P. Legrain. 2001. The protein-protein interaction map of Helicobacter pylori. Nature **409**:211–215.

Rubin, G., M. Yandell, J. Wortmann et al. (52 co-authors) 2000. Comparative Genomics of the Eukaryotes. Science **287**:2204–2215.

Sandberg, R., G. Winberg, C.-I. Bränden, A. Kaske, I. Ernberg, and J. Cöster. 2001. Capturing Whole-Genome Characteristics in Short Sequences Using a Naive Bayesian Classifier. Genome Res. **11**:1404–1409.

Saxonov, S., I. Daizadeh, A. Fedorov, and W. Gilbert. 2000. The Exon-Intron Database: An exhaustive database of protein-coding intron-containg genes. Nucl. Acids. Res. **28** (1):185–190.

Scala, A., L. Amaral, and M. Barthélémy. 2001. Small-world networks and the conformation space of a short lattice polymer chain. Europhys. Lett. **55** (4):594–599.

Schuster, P., W. Fontana, P. Stadler, and I. Hofacker. 1994. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. Proc. Roy. Soc. Lond. B **255**:279–284.

Schwikowski, B., P. Uetz, and S. Fields. 2000. A network of protein-protein interactions in yeast. Nature Biotechn. **18**:1257–1261.

Seidel, H., D. Pompliano, and J. Knowles. 1992. Exons as Microgenes. Science **257**:1489–1490.

Smith, C., J. Patton, and B. Nadal-Ginard. 1989. Alternative splicing in the control of gene expression. Annu. Rev. Genet. **23**:527–577.

Sprinzl, M., C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. 1998. Compilation of tRNA sequences and sequences of tRNA genes. Nucl. Acids Res. **26**:148–153.

Stanley, H., S. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng, and M. Simons. 1999. Scaling features of noncoding DNA. Physica A **273**:1–18.

Stoltzfus, A., D. Spencer, M. Zuker, J. Logsdon Jr., and W. Doolittle. 1994. Testing the Exon Theory of Genes: The Evidence from Protein Structure. Science **265**:202–207.

Tupler, R., G. Perini, and M. Green. 2001. Expressing the human genome. Nature **409**:832–833.

Uetz, P., L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshorn, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. Rothberg. 2000. A comprehensive analysis of protein-protein interactions of saccharomyces cerevisiae. Nature **403**:623–627.

Vazquez, A. 2001. Statistics of citation networks. http://arxiv.org/abs/cond-mat/0105031.

Venter, J., M. Adams, E. Myers et al. (271 co-authors) 2001. The sequence of the Human Genome. Science **291**:1304–1351.

Wagner, A., and D. Fell. 2001. The small world inside large metabolic networks. Proc. R. Soc. Lon. B **268**:1803–1810.

Waterman, M. 1995. Introduction to Computational Biology. Chapman and Hall, London.

Waterman, M. S., and T. F. Smith. 1978. RNA secondary structure: A complete mathematical analysis. Math. Biosci. **42**:257–266.

Watts, D. 1999. Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton University Press, Princeton, New Jersey, USA.

Watts, D., and S. Strogatz. 1998. Collective dynamics of 'small-world' networks. Nature **393**:440–442.

Wuchty, S. 2001. Scale-free Behavior in Protein Domain Networks. Mol. Biol. Evol. **18** (9):1694–1702.

Wuchty, S., W. Fontana, I. Hofacker, and P. Schuster. 1999. Complete Suboptimal Folding of RNA and the Stability of Secondary Structures. Biopolymers **49**:145–165.

Xenarios, I., E. Fernandez, L. Salwinski, X. Duan, M. Thompson, E. Marcotte, and D. Eisenberg. 2001. DIP: the Database of Interacting Proteins: 2001 update. Nucl. Acids Res. **29** (1):239–241.

Zipf, G. 1949. Human behaviour and the principle of least effort. Hafner, New York.

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. Science **244**:48–52.

Zuker, M. 2000. Calculating nucleic acid secondary structure. Curr. Opin. Struct. Biol. **10**:303–310.

ZUKER, M., and D. SANKOFF. 1984. RNA secondary structures and their prediction. Bull. Math. Biol. **46** (4):591–621.

ZUKER, M., and P. STIEGLER. 1981. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. Nucl. Acids Res. **9**:133–148.

APPENDIX B

Publications

Stefan Wuchty, *Scale-free Behavior in Protein Domain Networks*, Mol. Biol. Evol., **18**(9), 1694-1702, (2001)

Stefan Wuchty, *Interaction and Domain Networks of Yeast*, re-submitted after minor revisions to Proteomics

Stefan Wuchty, *The large scale organisation of genomic sequence segments*, submitted

Stefan Wuchty, *Small Worlds in RNA*, submitted

APPENDIX C

Curriculum vitae

Stefan Wuchty

* 11.6.1972, Wien

1978 – 1980	Volksschule Lingenau, 6951 Lingenau in Vorarlberg
1980 – 1982	Volksschule Kleistgasse, Kleistgasse 12, 1030 Wien
1982 – 1990	Bundesrealgymnasium BRG III, Radetzkystrasse 2a, 1030 Wien
5.90	Reifeprüfung
1990 – 1991	Präsenzdienst FMAR, Maria-Theresienkaserne Fasangartenstrasse 8, 1120 Wien
1991 – 1997	Studium der Chemie, Studienzweig Biochemie an der Universität Wien
9.94	1. Diplomprüfung Chemie
3.97 – 2.98	Diplomarbeit am Institut für Theoretische Biochemie an der Universität Wien
3.98	2. Diplomprüfung Chemie mit Auszeichnung
5.98 – 3.99	Wissenschaftlicher Mitarbeiter IPHT Jena, Merck KGaA, Darmstadt, Deutschland
4.99 – 7.02	Wissenschaftlicher Mitarbeiter European Media Laboratory (EML) Heidelberg, Deutschland

7.02	Promotion zum Doktor der Naturwissenschaften
8.02 –	Post-Doc, Department of Physics, University of Notre Dame, Indiana, USA