Topological Aspects of the Human Protein Interaction Network

Dissertation

zur Erlangung des Grades
"Doktor der Naturwissenschaften"

am Fachbereich Biologie
der Johannes Gutenberg-Universität
in Mainz

Thomas Wiebringhaus
geb. in Marl

Mainz, September 2008

Contents

1 Introduction **1**

1.1 Motivation, Background and Aim 2

1.2 Summary and Structure of the Thesis 3

2 Basic Properties **7**

2.1 Basic Properties 8

2.2 Clustering 10

2.3 Scale-free networks 11

2.4 Degree and Clustering Coefficient distribution 12

3 Intermodular Signatures **13**

3.1 Introduction 15

3.2 Methods 16

3.2.1 The Betweenness Centrality 16

3.2.2 The Pearson Correlation Coefficient 16

3.2.3 The InterConnectedness Coefficient 17

3.3 Results 18

3.3.1 High Betweenness- Low Connectivity 18

3.3.2 The InterConnectedness Coefficient 23

3.4 Discussion 26

3.4.1 High Betweenness- Low Connectivity 26

3.4.2 The InterConnectedness Coefficient 27

4 Biological Modules **29**

4.1 Introduction 31

4.2 Methods 31

4.2.1 The Clique Percolation Method CPM 31

4.2.2 The Network Clustering Coefficient 32

4.2.3 Biological Annotation of Communities 32

4.3 Results 32

4.3.1 Biological Annotation and Local Structure of Communities 32

4.3.2 Global Community Structure 34

4.4 Discussion 45

5 Topological Aspects of Signal Transduction **47**

5.1 Introduction 49

5.2 Methods 50

5.2.1 The skeleton or communication kernel 50

5.2.2 The degree-degree correlation 51

5.3 Results 51

5.3.1 Communities of Cell Communication 51

5.3.2 The skeleton or communication kernel 60

5.3.3 The degree-degree correlation 65

5.3.4 The *rich-club* phenomenon 66

5.4 Discussion and Outlook 68

5.4.1 Communities 68

5.4.2 The skeleton 70

5.4.3 Degree-correlation 71

Abstract **73**

Bibliography **79**

Supplementary **89**

Chapter 1

Introduction

1.1 Motivation, Background and Aim

Biological networks have evolved to carry out complex cellular behaviour using DNA, nucleic acids, proteins and molecules. Cells have to transpose and realize different functional and structural demands, e.g. transducing signals very reliable and fast from the plasma membrane to the nucleus, or maintaining biological defined structures, e.g. organelles and multiprotein complexes.
Depending on the task, very different topological architectures and networks have evolved. One goal of this thesis is the empirical examination of the topological features that mirror biological structures, functions and signaling networks; or in other words, this thesis aims to understand biological *network-to-function* relationships.

> "*But instead of a cell dominated by randomly colliding individual protein molecules, we now know that nearly every major process in a cell is carried out by assemblies of 10 or more protein molecules. And, as it carries out its biological functions, each of these protein assemblies interacts with several other large complexes of proteins*"
> Bruce Alberts [Alberts, 1998]

It is currently widely accepted that the understanding of complex cell functions depends on an integrated network theoretical approach and not on an isolated view of the different molecular agents [Hartwell et al., 1999]. A theoretical framework is provided when depicting a cellular network with graph theoretical methods that enables the use of the developed concepts in these mathematical fields.

Interestingly, organisational features among biological and between non-biological, e.g. technical or sociological networks are shared, so that unique emergence principles might exist [Barabasi, 2005]. Topological properties can be compared and advances in one field provide new advantages for other areas. Biological networks have a long evolutionary history and are thus of great importance for the understanding of complex networks at all.
Power-law distributions, the presence of scale-invariance and modularity are of no direct use for the biomedical experimentator, as these parameters are pure theoretical features. On the contrary, the examination of global characteristics give new insights not only into the organisational principles of modular networks but provide also a new view on specific cellular functions from a topological and pure mathematical perspective.
Current research has brought new possibilities to bring together conventionally separated fields as much as different as technology, sociology and biology as these networks have common organisation principles. Network biology is therefore highly interdisciplinary and can phenomenologically examined on three layers:

1. global graph theoretic view
2. hierarchical levels of regional subsystems
3. individual local network motifs

As the goal of this thesis is *from binary interactions to cellular biology*, mainly the regional subsystems are of interest.

The presented network is a partial human interactome network consisting of 9222 proteins and 36324 interactions. To examine a large interactome dataset, no isoforms and no cell-specific information is filtered (which is nevertheless only partially available). To study pure topological aspects, only binary interactions reliably extracted from peer-reviewed scientific publications, without secondary or associated information like biological function or subcellular localisation, is used here and hence the approach is almost free of a biological hypothesis.

1.2 Summary and Structure of the Thesis

Global characteristics of the presented human interactome network are shortly reviewed in chapter 2 and are examined in more detail by Schueler [Schüler, 2006].

> "*A functional* module *is, by definition, a discrete entity whose function is separable from those of other* modules." [Hartwell et al., 1999]

> *Are there topological properties or bottlenecks in a network theoretical manner that mirror biological structures or functions?*

In general there are two approaches, focusing on inter- or intramodular aspects. *Inter*modular links, proteins connecting *modules* representing biological functions, are assumed here to act as important key players in cellular biology and may give rise to adjacent functional subnetworks. But the actual *size* and protein composition of specific functions is unknown. From the notion of functional biology these links might be (rate-limiting) enzymes or whole pathways connecting a ligands activity to its cellular response including the whole complex expression of various genes necessary for the respective function. Regarding structural aspects, a *module* might also be a defined structural complex as the proteasome, responsible for protein degradation. Recently, the examination of intermodular links was approached by utilizing the *Betweenness Centrality* [Joy et al., 2005]. This method determines proteins highly frequent used for shortest paths in the context of the whole protein network. Chapter 3 compares the published findings for yeast to the presented human PPI network. Similar results were obtained by determining the distribution patterns for the *Betweenness Centrality* while the human network also establishes *High Betweenness and*

Low Connectivity proteins which are biologically analysed and interpreted here as potential intermodular links, or more specifically, as shuttling proteins between organelles.
As an optimisation for finding modular links, a new method is developed here based on proteins located between highly *clustered* regions potentially mimicking biological interrelationships much better, than regarding highly *connected* regions (proposed by [Hwang et al., 2006]). As a proof of principle, the "Mediator complex" is found in first place, the prime example for a connector complex, validating the approach.

Complex real networks are organised into *modules* [Hartwell et al., 1999], [Ravasz & Barabási, 2003].

> *What does the modular nature mean in detail?*
> *Are biological structures or functions topologically discriminable?*
> *Do these* modules *represent biological aspects?*

Focusing on *intra*modular aspects, the measurement of *k-clique communities* (also known as *Clique Percolation Method CPM*) discriminates overlapping cliques and *communities* very well as shown for different hierarchical networks, but not for human PPI networks so far [Derényi et al., 2005], [Palla et al., 2005], [Adamcsek et al., 2006]. Chapter 4 analyses and functionally interprets ∼20 of the largest *k-clique communities* found in the presented human network in detail, like multiprotein complexes, e.g. the transcription machinery, the proteasome and the actin-related complex, or smaller functional complexes, e.g. the NF-κB complex. Very interestingly, two highly interconnected and relatively large subgraphs for signal transducer and transcription factor proteins are found, an observation which is further processed in chapter 5. Statistical properties of the *community* structure reveal a similar observation for unique network properties found by Palla et al. also for the presented human network.

> *How can regulatory complexity and diversity of signal transduction networks or transcription events be explained with only a limited number of proteins?*
> *What is the use that some specific proteins have more than 100 interaction partners and some only a few? And what is the molecular and biological nature of these proteins with high interactivity?*
> *Which proteins are most central for communication and is this property mirrored by an underlying structure with topological constraints?*

Chapter 5 is at the heart of the thesis and examines the large *communities* for signal transduction and transcription found by the method of *k-clique communities* (chapter 4) in detail. It is revealed that the two subnetworks are highly structured and networked, allowing manifold regulatory events for transducing a signal from the cytoplasm to the nucleus as well as enabling a high diversity of

transcription events in the nucleus. The core of this signaling *module* consists of proteins very central for cell growth, mitosis and differentiation and is enriched with protein domains, especially SH2 and SH3 domains. This core might represent a central component for integrating diverse signals and putatively regulate pathway *cross-talk*. Interestingly, a small functional bottleneck for transducing a signal from the cytoplasm into the nucleus and thus into the *community* for transcription is found by pure topological properties.
To understand the organisation of cellular communication processes better, the *skeleton* or communication kernel is analysed as well, based on interactions which are highly transferred from an information flow aspect. Very interestingly, it is observed that proteins highly interwired and thus yielding the highest number of *shortcuts* in the original network, reflect signaling or transcription proteins. The high number of *shortcuts* represent regulatory diversity and maximally secured information exchange as well. Moreover the most interconnected proteins show a high overlap to the central signaling or transcription cores recovered and explored in chapter 4 and 5 by applying an independent graph theoretical method here. The presented human *skeleton* is globally also scale-free, firstly shown for a human PPI network and exhibits the longer-loop dominant structure, a notion which is not further studied and interpreted here but will yield a first basis for studying self-similar properties of biological networks ([Song et al., 2005], [Song et al., 2006]; see also outlook). It is also shown that most of the human top hub proteins either constitute a *rich-club* [Colliza et al., 2006] for signal transduction or for transcription factors. Although *rich-clubs* are present, the degree-degree correlation is dissortative (repulsive) which exhibits an interesting and unforeseen property of a complex network.

Chapter 2

Basic Properties

2.1 Basic Properties

A few graph-theoretical notions are given or mentioned here, for more details refer to [Barabási & Oltvai, 2004] and references therein, or to a typical text book about graph theory.
PPI-networks are usually represented as graphs. A graph *G=G(V,E)* is a set of objects called *vertices* or *nodes V* joined by links called *edges E*. In this study the *nodes* represent the proteins while the *edges* represent direct physical interactions between the proteins. *Vertices* that are joined by an *edge* are called adjacent or *neighbours*. The *degree d(v)* of a *vertex v* is the number of *neighbours* or interactors of *v*. The average *degree* $<d(v)>$ of an undirected graph G is thus given by:

$$< d(v) > = \frac{2E}{V} \tag{2.1}$$

Note that *d(v)* is also denoted as *k*.
A *shortest path* between two *nodes u* and *v* is a path connecting *u* and *v* with the minimal possible number of *edges* and is denoted by *p(u,v)*. The *diameter* is the pathlength of the longest shortest path in the network. The *mean path length* $<l>$ is given by the sum of the lengths of all shortest paths in a network divided by the number of all shortest paths in the network. Large networks featuring a small *diameter* as well as a small *mean path length* are called *small-world networks* [Watts & Strogatz, 1998]. A graph is called *connected* if a path between every pair of *nodes* in G is possible and *disconnected* otherwise.
Three eukaryotic protein networks were compiled by Schueler [Schüler, 2006] and basic properties were compared to recently published networks in the literature (see tables 2.1, 2.2). As the *Human Interactome Map (HIM)* is used in this study as well, a few network properties are shortly reviewed here and partially compared to other networks.

Table 2.1: Basic properties of the curated eukaryotic interactome networks (taken from Schueler, 2006). The compiled interactome networks were designated as HIM for "Human Interactome Map", FIM ("Fly Interactome Map") and YIM ("Yeast Interactome Map")

Homo sapiens	HIM	[Stelzl et al., 2005]
Proteins	9,475	1,713
Interactions	36,495	3,150
components	111	46
main component	9,225 (97.36%)	1,604 (93.6%)
$<d(v)>$	7.7	1.84

Drosophila melanogaster	FIM	[Giot et al., 2003]
Proteins	7,499	4,651
Interactions	24,991	3,039
Components	60	591
main component	7,372 (98.3%)	3,039 (65%)
$<d(v)>$	6.66	2.04

Sacharomices cerevisiae	YIM	[Ito et al., 2001] and [Uetz et al., 2000]
Proteins	5,298	1,417
Interactions	50,434	1,520
Components	3	160
main component	5,294 (99.9%)	970 (68%)
$<d(v)>$	19.04	2.15

Table 2.2: Mean path length of three eukaryotic networks (taken from [Schüler, 2006])

Network	$< l >_{\text{observed}}$	$< l >_{random}$
HIM	3,959	4.486
FIM	4,222	4.706
YIM	3,313	2.91

2.2 Clustering

A measurement for the local clustering of a *vertex* is the *clustering coefficient CC1(v)*. The *clustering coefficient* is defined by the number of *edges* between all *neighbours* of v divided by the maximal number of *edges* between the *neighbours* of v.

$$CC1(v) = \frac{2E(v)}{d(v) \cdot (d(v) - 1)} \tag{2.2}$$

The average *clustering coefficient* $<CC1(v)>$ of a network is the sum of the *clustering coefficients* of each *vertex* divided by the number of *vertices*.

$$< CC1(v) >= \frac{1}{V} \sum_{v \in G} CC1(v) \tag{2.3}$$

Table 2.3 shows some empirical results for various networks and table 2.4 for eukaryotic networks. It is shown that the explored parameters are larger that expected by chance, demonstrating that clustering is a unique feature also for eukaryotic protein interaction networks.

Table 2.3: Empirical results for the degree of clustering in large networks. N = number of *nodes*; $<d(v)>$ = average degree; $<CC1(v)>$ = observed average clustering coefficient; $CC1(v)_{rand}$ = expected clustering coefficient for a random graph with N *nodes* (taken from [Newman, 2002], [Przulj, 2005])

	N	$<d(v)>$	$<CC1(v)>$	$< CC1(v) >_{rand}$
Internet (autonomous system)	6,374	3.8	0.24	0.0006
World Wide Web	153,127	35.2	0.11	0.00023
Power Grid	4,941	2.7	0.08	0.00054
Biology Collaborations	1,520,251	15.5	0.081	0.00001
Mathematics Collaborations	253,339	3.9	0.15	0.000015
Film-actor Collaborations	449,913	113.4	0.2	0.00025
Company Directors	7,673	14.4	0.59	0.0019
Word Co-occurence	460,902	70.1	0.44	0.00015
Neural Network of C. elegans	282	14	0.28	0.049
Metabolic Network	315	28.3	0.59	0.09
Food Web	134	8.7	0.22	0.065

Table 2.4: Average clustering coefficients for eukaryotic protein networks (taken from [Schüler, 2006])

	$<CC1(v)>$	$< CC1(v) >_{rand}$
HIM	0.114	∼0.0008
Stelzl.	0.006	∼0.0021
FIM	0.029	∼0.0008
Giot	0.11	∼0.0004
YIM	0.154	∼0.0035
Ito-Uetz	0.09	∼0.0015

2.3 Scale-free networks

Many real complex networks exhibit some *nodes* with a significantly higher connectivity as average. Networks that feature this property are called scale-free networks (see fig. 2.1).

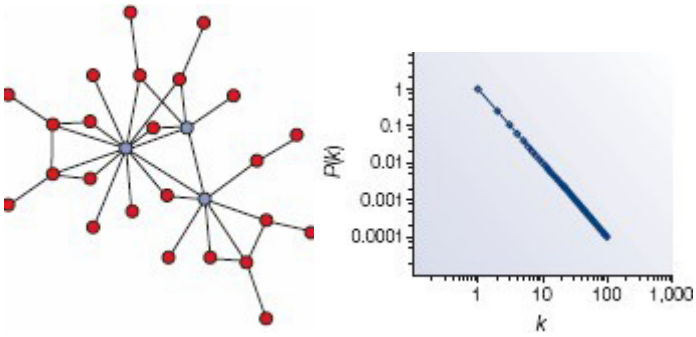


Figure 2.1: A scale-free network and its degree distribution

The *degree* distribution of a scale-free network can be approximated by a power law:

$$P(k) \sim k^{-\gamma} \tag{2.4}$$

Note that the *degree* k is the same as *d(v)*. The *degree* exponent γ indicates how distinct the scale-free property of a given network is. The unusual properties of a scale-free network emerge for $2 < \gamma < 3$, for γ values smaller than 2 a *hub-and-spoke network* emerges in which the most connected *node* is connected to a very large fraction of the whole network and networks with $\gamma >$

3 show a behaviour very similar to random networks. The mechanism how scale-free networks evolve is called *growth* and *preferential attachment* and implies that the probability of a *node* getting a new neighbour increases with the *nodes* [Barabási & Oltvai, 2004](also known as *rich-get-richer principle*). Another very interesting property associated with scale-free networks is their robustness against random elimination of *nodes*. Up to 80% of the networks *nodes* have to be randomly eliminated to disconnect a connected scale-free network. However, this robustness comes at the cost of a high vulnerability against a targeted elimination of the most highly connected *nodes*.

2.4 Degree and Clustering Coefficient distribution

The *degree* distribution *P(k)* yields the probability that a selected node has exactly k links. The *degree* distribution discriminates between different types of networks. A peaked *degree* distribution indicates a network in which all *nodes* have very similar degrees while a slowly decaying *degree* distribution indicates that the network is held together by a few highly connected *nodes*, commonly referred to as *hubs*.
The distribution of the *clustering coefficient CC1(v)* is a magnitude which measures the networks tendency to build hierarchical clusters. Networks that are hierarchically organised show an interdependency between a *nodes degree* and its clustering coefficient, a high degree correlates with a low clustering coefficient in these networks [Barabási & Oltvai, 2004]. The presented and analysed human network is hierarchically organised (see table 2.5 and [Schüler, 2006]).

Table 2.5: Power-law exponent β for the distribution of the clustering coefficient as a function of the degree in the HIM, FIM and YIM network (taken from[Schüler, 2006])

	HIM	FIM	YIM
β	-0,5763	-0,4409	-0,3413

Chapter 3

Intermodular Signatures

Abstract

Protein interactions are organised into structural and functional modules to perform complex cellular behaviour. The identification of topological properties for proteins connecting or mediating functions will provide important new insights into the organisation of biological network structures. It is suggested here that inter-organelle communication, especially intracellular transport mechanisms, are topologically distinguishable in a homo sapiens protein interaction network by calculating low-connected proteins lying frequently on geodesic paths.
In addition, a new graph theoretical parameter is introduced to find proteins located between densely clustered regions by regarding regional information. As a proof of principle, the well-known biological connector complex *mediator* for transcription initiation is found. The findings propose that some essential cellular properties like organelle communication or assemblies of transcription complexes are present as structural signatures in human cells.

3.1 Introduction

One major question in current network biology is the decomposition of cellular behaviour into smaller biological functions to understand the complex interwoven web of protein interactions. Cellular tasks are carried out by a number of different proteins and vice versa one protein can take part in many different functions [Gavin et al., 2002]. It is a well-known condition that functions or *modules* have to be connected to exchange information and thus have to be linked via intermodular links or mediating complexes.
Full-connected complexes are regarded as molecular multiprotein complexes, but not any biological function is carried out by dense subnetworks. By applying established methods from graph theory, a new interesting topological feature was recently discovered for yeast protein interaction networks by calculating the *Betweenness Centrality BC* [Joy et al., 2005]. The *BC* measures how central a specific network component is for the overall communication of a network. Although this attempt is only technically meaningful here, as the calculated shortest paths are not necessarily biological pathways, the method might be practicable for the calculation of global bottlenecks.
It is expected that the *BC* is proportional to the *degree* because more information is capable to flow over proteins that have more interaction partners as shown recently [Nakao, 1990], [Goh et al., 2003]. In other words, information can be spread more slightly over a scale-free network, a notion which is also manifested by the presence of hubs. This proportionality holds also true in particular for yeast proteins having more than about 10 interaction partners, as estimated here for the *S. cerevisiae* data in the Joy et al. paper. In the opposite, the distribution of the *BC* regarding *nodes* with less than 10 interaction partners is much more dispersed. The authors suggested proteins having *high betweenness* and relative *low connectivity* (*HBLCs*) as putative interfunctional connectors in yeast [Joy et al., 2005].
This chapter describes the first attempt to examine whether the findings for *S. cerevisiae* are also present for *Homo sapiens*. Furthermore, in contrast to Joy et al., it is deeply analysed if these *HBLCs do* represent intermodular links by interpreting the results from a biological viewpoint.

Biological networks have evolved to fulfill different cellular tasks and so their structural features require different methodological aspects to comprise. Recently, the *bridging centrality* was developed for finding intermodular connections by establishing a combined measurement of the *Betweenness Centrality BC* and the *bridging coefficient*, which measures how well a *node* is located between highly *connected nodes*, and thus *global* and *local* parameters are combined [Hwang et al., 2006]. While the *BC* finds bottlenecks based on information flow, the *Bridging Coefficient* adds local topological parameters by means of *node* connectivity. Visual interpretation shows in their study that the parameter finds bridging proteins in five real networks coming from biology, technology and sociology. However, the largest network comprises only 359 *nodes* and 435 *edges*, representing rather small networks in their study.

Preliminary results coming from the application of the *bridging centrality* to the presented much larger human network (factor ∼30 larger) shows that mainly proteins located between hubs represent high *bridging centrality* values (data not shown), as expectable regarding the formulation of the parameter. However, only some (date-) hubs might represent a functional *module* [Han et al., 2004]. In this study solely *regional* parameters should be considered for finding inter-modular connections. As a new approach, a method called *InterConnectedness Coefficient IC* based on proteins located between highly *clustered* and not highly *connected* regions is introduced. Detailed manual curation is performed for the top scoring proteins.
The methods section describes the notion of *HBLCs* and the constraints resulting in the development of the *IC*. The results section shows and compares the distribution of the *BC*, examines the molecular basis of *HBLC nodes* and presents some manually curated high ranking *nodes* from a biological viewpoint. The results from the application of the *IC* are presented and afterwards outstanding biological examples of the feasibility of the *IC* are shown as well. The last section summarizes and discusses the presented results and points out putative further directions and developments for the identification of modular links.

3.2 Methods

3.2.1 The Betweenness Centrality

The *Betweenness Centrality BC(v)* measures the frequency of a *node* for lying on a geodesic (shortest) path:

$$BC(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.1}$$

with σ_{st} the number of shortest paths between s and t with $\sigma_{st}(v)$ the number for *node* v lying on shortest paths between s and t.
Calculation is done with the network software Pajek [Batagelj & Mrvar, 1998].

3.2.2 The Pearson Correlation Coefficient

The Pearson Correlation Coefficient $r_{\mathrm{x,y}}$ is applied to the function of *BC(v)* and *d(v)* to calculate a linear correlation:

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2}$$

with x_i = BC(v), y_i = *d(v)*, the averages $\bar{x}$ and $\bar{y}$ and n the number of data [Kim et al., 2004].

3.2.3 The InterConnectedness Coefficient

A new approach for finding intermodular links is proposed and developed following ideas from a biological perspective. Dense non-overlapping distinct regions representing functions or *modules* may be connected via less dense proteins or regions and thus inter-modular aspects are examined here.
The basic idea of the *IC* is that *regional* characteristics rather than *local* and/or *global* properties are regarded. Shortly, the *IC* describes the *clustering* of a *node* related to the average *clustering* of the direct neighbours as a node weight and thus should yield (at most non-clustered) proteins connecting dense regions.

The *Clustering Coefficient* is

$$CC1(v) = \frac{2l_i}{d(v) \cdot (d(v) - 1)} \tag{3.3}$$

$0 \leq CC1(v) \leq 1$

with *d(v)* the *degree* of *node* v and l_i the number of links among interactors of v.

To ensure a *repulsion* between *modules*, *node* v should be minimally clustered with direct neighbours and hence a simple *repulsion coefficient* is calculated for *node* v:

$$R(v) = 1 - CC1(v) \tag{3.4}$$

$0 \leq R(v) \leq 1$

Incident *edges* linking the connector to the *module* should not be implied in the average *clustering coefficient* of neighbours and thus the denominator of the *CC1(v)* is modified into the *CCi*. This modification ensures that the *clustering* of the putative adjacent *module* is still maximal (equal to 1) when one single edge (the connector) traverses into/ out of the *module*. This yields in a favourisation of an adjacency to complete subgraphs.

$$CC_i(v) = \frac{2l_j}{(d(v)-1)(d(v)-2)} \tag{3.5}$$

with l_j as the number of links among interactors of i
Thus the *IC(v)* is given by

$$IC(v) = R(v)\frac{1}{d(v)}\sum CC_i(v) \tag{3.6}$$

3.3 Results

3.3.1 High Betweenness- Low Connectivity

Regarding the *Betweenness Centrality* as a function of the *degree*, a similar pattern and correlation is found comparing yeast [Joy et al., 2005] to human (see table 3.1). The *Pearson Correlation Coefficient* of the *degree d(v)* and *BC(v)* clearly establishes a linear correlation (see figure 3.1 and table 3.2), disregarding proteins with *degree* one having a *BC(v)* of zero. This correlation is also in accordance with two recently published surveys about social networks [Nakao, 1990], [Goh et al., 2003].
However, regarding the distribution pattern in figure 3.1 a profound variation is apparent for proteins with approximately $d(v) < 11$. Comparing the results for yeast [Joy et al., 2005] and human (this study), the range of the *BC(v)* is two orders of magnitude larger and the scattering of *BC(v)*, especially for lower degrees, is more pronounced in the presented human network (figure 3.1, table 3.1, table 3.2).

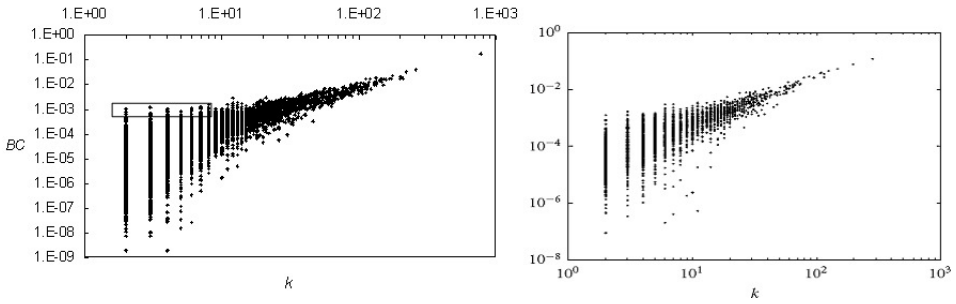
Table 3.1: Number of *nodes*, *edges* and range of the *betweenness centrality* for the two networks. Data for *S. cerevisiae* is taken from [Joy et al., 2005] and the range of the *BC(v)* is estimated from their published figure.

species	# of *nodes*	# of *edges*	range of *BC(v)*
S. cerevisiae	2605	6438	10^{-7} to 10^{-1}
H. sapiens	9222	36324	10^{-9} to 10^{-1}

Although the low connectivity of *HBLCs* would imply a less important role for general network function, the high *BC(v)* values denote an important role with a global impact. As much as 5000 proteins have a *degree* less than 11 and a *BC(v)* differing in the range of 10^{-9} to 10^{-2} (see table 3.2).

Table 3.2: The Pearson correlation $r_{x,y}$ for different degree ranges of the human network. The correlation between *BC(v)* and *nodes degree* for proteins with less than 11 interactors is much lower than the correlation for proteins with *degree* larger than 10.

degree d(v)	# of *nodes* V	$r_{x,y}$
>1	6977	0,85
>10	1798	0,89
>1 and <11	5179	0,45



a b

Figure 3.1: Degree-BC correlation (d(v) denoted as k) plotted in logarithmic scale a) this study human and b) figure taken from Joy et al. for yeast. The rectangle denotes HBLCs in a)

Biological interpretation of HBLCs

To examine if *HBLCs* represent biological relevant intermodular connections, the molecular basis of the top 20 proteins is resolved and interpreted biologically. As anticipated, some proteins being part of full-connected complexes have a *BC* of zero, indicating that the information flows over the complex and not over single *nodes* (see table 3.3).

Table 3.3: Tata-binding protein associated factor TAF8 and breast cancer metastasis-suppressor 1-like BRMS1L exhibit a *degree* less than 11 and a *BC(v)* of zero. Both proteins belong to large subcomplexes and thus *shortest paths* central for the *BC(v)* calculations do not *flow* over these *nodes*.

	d(v)	BC(v)	complex
TBN (TAF8)	10	0	TAF
BRMS1L	9	0	Sina3 and HDAC

Very interestingly, proteins important for protein shuttling between organelles are found in first place, indicating a biological meaningful basis. Import and transport mechanisms comprise peroxisomal import, nucleus import or ER-to Golgi transport (see table 3.4).

Table 3.4: Subset of the top 20 proteins with $d(v)<11$. Biological functions and trafficking and transport types are annotated by examining the literature (see text).

Symbol	biological function	*d(v)*	*BC(v)*	trafficking	transport type
CGI-37 (NIP7)	nuclear pore	2	1,03E-03	cytoplasm -> nucleus	gated
SEC61B	translocator	6	9,24E-04	cytoplasm -> ER	trans-membrane
PEX14	peroxisomal import	8	1,57E-03	cytoplasm -> peroxisome	vesicular
COG5	membrane trafficking	4	1,03E-03	intra-Golgi retrograde trafficking	vesicular
DCTN2		9	1,02E-03	ER -> Golgi	vesicular
GOLGB1	vesicular transport	10	8,09E-04	ER -> Golgi	vesicular
NOL8		4	8,22E-04	nucleolus	
ABCA1		9	8,82E-04	membrane/ cholesterol transport	
IL10RB	receptor	8	1,04E-03	extracellular -> membrane -> cytoplasm	
CD59	receptor	10	9,01E-04	extracellular -> membrane -> cytoplasm	
MC4R	receptor	8	1,05E-03	extracellular -> membrane -> cytoplasm	

Gated transport through the Nuclear Pore Complex NPC

The nuclear pore complex (NPC) is the *sole* gateway between the nucleus and the cytoplasm of eukaryotic cells. It mediates *all* trafficking between these 2 cellular compartments. Nuclear import homolog NIP7 (CGI-37) interacts with NOL8 and takes part in the RCC1-Ran pathway [Sekiguchi et al., 2004]. The nuclear import of proteins through nuclear pore complexes is an intricate organised set of interactions between cargoes, carriers where Ran GTPase and

RCC1 act as molecular switches ([Stewart, 2007] and fig. 3.2).

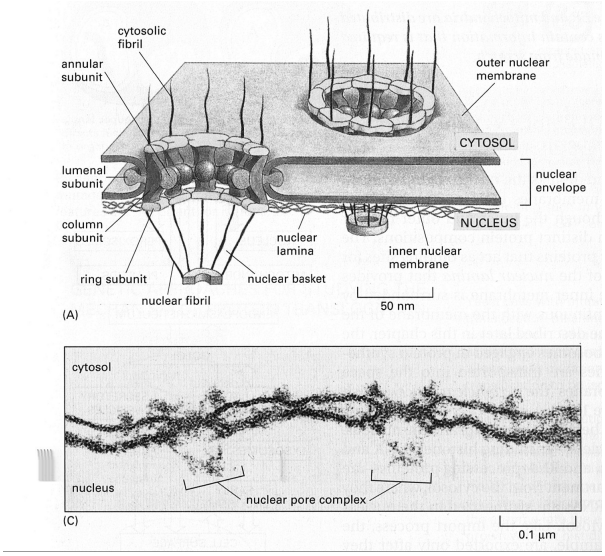


Figure 3.2: The Nuclear Pore Complex NPC (taken from [Alberts, 2002]) The NPC is the sole gateway between the cytoplasm and the nucleus.

The SEC61 complex forms the protein translocator of the ER

The Sec61 complex is the central component of the protein translocation apparatus of the endoplasmic reticulum (ER) membrane. Oligomers of the Sec61 complex form a transmembrane channel where proteins are translocated across and integrated into the ER membrane (adapted and taken from NCBI Gene ID: 10952)(see fig. 3.3).

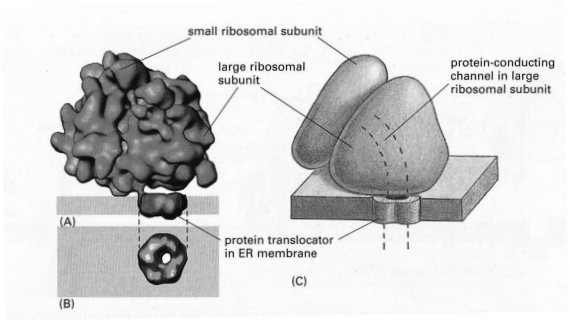


Figure 3.3: The protein translocator of the ER is a well known transmembrane channel across ER membranes (taken from [Alberts, 2002])

The peroxisomal import machinery

The peroxisomal biogenesis factor 14 (peroxin) (PEX14) is an *essential* component of the peroxisomal import machinery. The protein is integrated into peroxisome membranes with its C-terminus exposed to the cytosol, and interacts with the cytosolic receptor for proteins containing a PTS1 peroxisomal targeting signal. The protein also functions as a transcriptional corepressor. (taken from NCBI Gene ID: 5195)

COG5 is part of the conserved oligomeric Golgi (COG) complex (Golgi-associated membrane trafficking)

The COG5 protein is part of the conserved oligomeric Golgi (COG) tethering complex which is *essential* for maintaining the structure and function of the Golgi apparatus. Protein trafficking along the secretory and endocytic pathway is primarily mediated by shuttling vesicles. The efficient and precise fusion of vesicles with the target compartment is thought to be achieved by bringing the vesicles to a close proximity with the receiving compartment via a process referred to as tethering [Loh & Hong, 2004], [Ungar et al., 2005].

Transport from the ER to the cis/medial Golgi compartments requires the action of VDP, GM130 and giantin (GOLGB1)

GOLGB1, also known as giantin, interacts with VDP p115 vesicle docking protein VDP. The protein encoded by this gene is a peripheral membrane protein which recycles between the cytosol and the Golgi apparatus during interphase. It is regulated by phosphorylation: dephosphorylated protein associates with the Golgi membrane and dissociates from the membrane upon phosphorylation. Transport from the ER to the cis/medial Golgi compartments requires the action of this gene product, GM130 and giantin in a sequential manner. (NCBI Gene GeneID: 8615)

Figure 3.4 depicts the major trafficking channels between organelles, *HBLCs* represent some of them as intermodular bottlenecks.

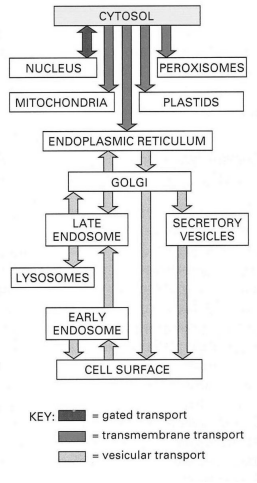


Figure 3.4: *HBLC* proteins examined here represent *nodes* of the three known transport systems in eukaryotic cells. (taken from [Alberts, 2002])

3.3.2 The InterConnectedness Coefficient

For a first analysis, the top 10 percentile of IC proteins having the highest *degree* are examined and biologically interpreted (see table 3.5).

Table 3.5: Subset of the top 10 percentile of IC proteins having the highest *degree*. All proteins are associated to structural multiprotein complexes important for transcription initiation.

protein	*d(v)*	*CC1(v)*	*1-CC1(v)*	*<CCi>*	*IntCoeff*	complex
IXL	25	0,247	0,753	0,421	0,317	Mediator
EG1(MED28)	24	0,250	0,750	0,456	0,342	Mediator
MED9	27	0,185	0,815	0,373	0,304	Mediator
ORC2L	22	0,355	0,645	0,292	0,188	Origin Recognition

Seven proteins among the top 16 facilitate Tata-binding protein (TBP)- associated factors (TAFs), a well-known connector complex. Proteins of the Spliceosome (SNRPD3 and SNRPE) and the Exosome Complex (SKIV2L2, EXOSC2 and EXOSC) are found as well, two cooperating complexes important for RNA manipulation. Proteins of the mRNA polymerase (esp. POLR2D and POLR2H) are also found, the complex associates via the mediator complex (IXL, EG1, MED9) with TFIID (TBPs and TAFs) and co-activating factors and initiates transcription of target genes. Further important revealed proteins are from the

Replication Factor/ DNA elongation (RFC4) and the Histone Deacetylase complexes (BRMS1 and RBP1). Figure 3.5 shows the distribution of the IC and the *degree*.

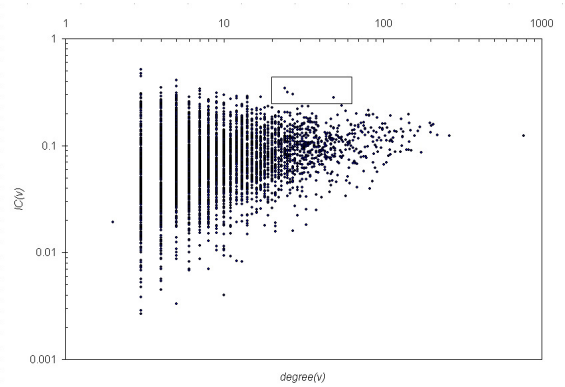


Figure 3.5: Degree- IC distribution. The rectangle denotes three proteins forming the mediator complex(left) and ORC2L of the origin recognition complex (right).

Biological interpretation of the IC

The mediator complex

The three proteins with the highest *degree* among the top 10 percentile are IXL, MED28 (EG1) and MED9, each is part of the mediator complex, an evolutionary conserved multisubunit protein complex that regulates the transcription process by RNA polymerase II. The Mediator complex acts as a molecular bridge between DNA-bound transcriptional activators and the transcription machinery ([Belakavadi & Fondell, 2006], see fig. 3.6 and 3.7).

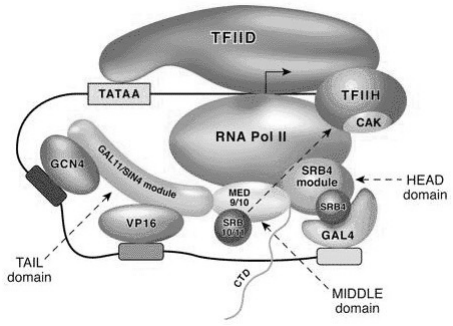


Figure 3.6: Illustration of the mediator complex (MED9/10) in yeast

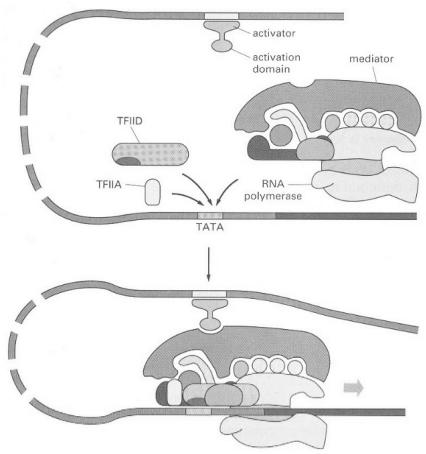


Figure 3.7: Binding of the mediator complex

The Origin Recognition Complex (ORC)

The protein encoded by this gene (ORC2L) is a subunit of the ORC complex. The origin recognition complex (ORC) is a highly conserved six subunits protein complex *essential* for the initiation of the DNA replication in eukaryotic cells. Studies in yeast demonstrated that ORC binds specifically to origins of replication and serves as a platform for the *assembly* of additional initiation factors such as Cdc6 and Mcm proteins. (taken from NCBI Gene ID: 4999, see fig. 3.8)

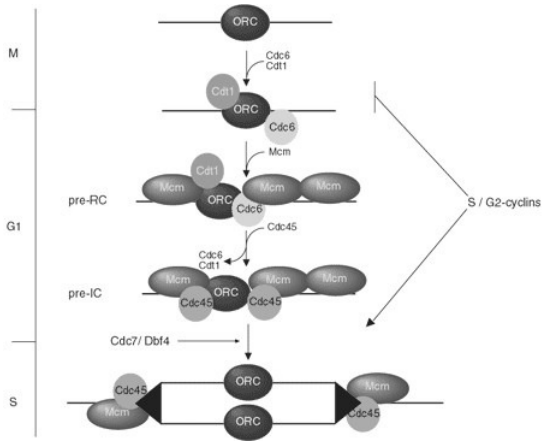


Figure 3.8: Illustration of orc assembly.

3.4 Discussion

3.4.1 High Betweenness- Low Connectivity

To summerise, the presence of *HBLCs* is examined here for a large human network. It is found that the distribution pattern of the parameter regarding *low connectivity* proteins is very similar to yeast and demonstrates that this topological feature is also evident for the examined human network. This finding is not explained by typical scale-free properties of protein interaction networks, where *low-connectivity* proteins also have *low betweenness centrality* [Nakao, 1990], [Goh et al., 2003]. It was supposed by Joy et al. that these *HBLCs* might represent functionally important proteins and act as intermodular connections.
In contrast to previous results for yeast, *high betweenness- low connectivity* proteins are biologically interpreted in detail here. It is shown that many of the top *HBLCs* represent biological meaningful bottlenecks for functional and structural compartmentalization processes. To be specific, proteins having a relative *high betweenness centrality* and *low degree* represent transport mechanisms from the ER to Golgi, the ER protein translocator, peroxisomal import or the nuclear pore complex.
Although the calculation of the *betweenness centrality* on the whole network is biologically not meaningful (a protein does not communicate to *any* other protein), the findings propose that from a technical perspective the *HBLCs* might represent structural narrow positions in the network. Related to their *degree*, an unproportional amount of information is exchanged over these *nodes* and *edges*.
Although the interpretation and validation is very complicated and time-consuming, the hypothesis that *HBLCs* might be intermodular links
[Joy et al., 2005] is in good agreement with the presented results here. The results suggest that subcellular compartmentalization barriers are mirrored by topological features. It is to mention that proteins with a low *degree* might also be less examined, although the depicted examples show a very specific topological signature. This study emphasis on the top percentile *HBLC* proteins (*degree* <11), further research how to measure a *degree* threshold would be advantageous.
The results suggest biological compartments as *modules*, as a separated location for specialized metabolism. Protein trafficking processes, which are observed here, represent communication among organelles. Interestingly, in addition to structural bottlenecks, some functional bottlenecks are also present: pyrin/marenostrin (MEFV) protein interaction with two cytosolic protein adaptors ASC (also called PyCARD) and PSTPIP functionally connect three important cellular pathways: apoptosis, cytoskeletal signaling and cytokine secretion [Schaner & Gumucio, 2005]. Also G-Protein coupled mechanisms are observed, reflecting a similar functional bottleneck. However, further research and interpretational effort is needed here to get to a deeper understanding.
To conclude, the topological feature of *HBLCs* is also present in the examined human network and (at least) one molecular basis might reflect organelle

import/ export mechanisms.

3.4.2 The InterConnectedness Coefficient

The *InterConnectedness Coefficient IC* is motivated by finding proteins located properly between *clustered* regions. The results suggest that the parameter detects *nodes* between highly interwired complexes. As the evaluation of the huge amount of results provides a wealth of data to interpret, the presented examples focus on proteins with a large *degree* as a first step. Interestingly, the examined examples underline that some *modules* are linked via other mediating modules. It is not conceivable that only one single protein constitutes the molecular bridge between *modules*, rather many proteins accomplish the communication between *modules*. To ensure proper functioning between modules, these proteins might be highly redundant by topology. This assumption is further supported by the preliminary findings that a *node cut* approach, deleting the top percentage of the IC nodes in an iterative manner, diminishes modularity as measured by the average *CC1(v)*, proposing that some larger *modules* are constituted of smaller *modules*. As a further optimisation, the parameter could be combined with the BC, although the BC range is much greater than the IC range. However, more research is needed to incorporate other parameters or to adjust the IC to favour other topological features. Note that both methods are virtually hypothesis-free as only topological characteristics are considered.

In conclusion, different biological characteristics employ distinct structural features and therefore the elucidation of the same demands different theoretical attempts. Future research on both, the coherence constraints of *modules* as well as the interfunctional restrictions will give a multitude of new interesting insights into the complex organisation of cellular behaviour.

Chapter 4

Biological Modules

Abstract

Cellular functions are performed by groups of functionally associated proteins also known as *modules* or *communities*. While some evolutionary highly conserved structural functions are carried out by densely packed multiprotein complexes as full-connected subgraphs, most of the *modules* are less compact and delimited and consist of highly overlapping smaller functional groups. The Clique Percolation Method CPM is shown to uncover dense interconnected units in various complex networks from life and social sciences. The method is firstly applied here to a large human PPI network. Numerous *communities* are found, biologically annotated and interpreted as known functional associations. Moreover, complete coherent sub-processes are reconstituted by pure topological aspects. The network of *communities* is analysed and the human *community* structure and its distribution characteristics are examined.

4.1 Introduction

The investigation of biological complexes and functions as *modules* is an active area of research [Luo et al., 2007], [Chen & Yuan, 2006]. In a protein interaction network, composed of binary protein interactions, functionally connected proteins appear as groups of densely interconnected nodes (proteins). However, the required density of connectedness necessary for a biological function is not known and will presumably vary among different functions. Finding these highly interconnected groups *a priori* will give new insights into the global organisation of such networks as well as new hypothesis for the mathematical constraints of a function. Moreover pure topological observations may predict and discover yet unknown biological associations.
Complex networks are hierarchically organised so that the presence of *communities* is a signature of real networks. Palla et al. examined the *community* structure for different complex networks in nature and society. Recently, the *Clique Percolation Method CPM* was also successfully applied to a PPI network of *S. cerevisiae* [Adamcsek et al., 2006]. This chapter examines the feasibility of finding dense multiprotein subcomplexes and/ or functionally active *modules* by calculating *k-clique communities* for a large human PPI network. Found *communities* are annotated in detail and the *community* structure is presented and interpreted. Moreover the scaling properties are analysed and compared to other complex networks recently described. The methods section describes the notion of *k-core communities* and introduces a new parameter for measuring network density, called the *Clustering Coefficient of Networks* CC_n. The results section presents more than 15 large *communities* annotated to well-known biological structures and functions. Additionally, the *community* structure is shown and analysed firstly for a human network. The discussion section interprets the results and compares the findings to the literature.

4.2 Methods

4.2.1 The Clique Percolation Method CPM

The method of *k-clique communities* [Derényi et al., 2005], [Palla et al., 2005], [Adamcsek et al., 2006], also known as *Clique Percolation Method (CPM)*,
is briefly described in the following. For algorithmic and other details see Palla et al. and their supplementary information. A *k-clique community* is composed of a group of functionally connected proteins where members are highly interconnected. The definition of a *community* represents the notion that its members can be reached through highly- or full- connected subsets of *nodes*, that means complete subgraphs that share their *nodes*. Or in other words, a *community* is the union of all full-connected subgraphs (*k-cliques*), that are reachable through adjacent *k-cliques*, and two *k-cliques* are adjacent if they share *k-1 nodes*.
To analyse and compare the *community* structure to other published networks,

four basic parameters introduced by Palla et al. are calculated here as well. The size *s(a)* of any *community a* is defined as the number of its *nodes*. Each *node v* of the PPI network can be characterised by a membership number m_v, which is the number of *communities* the *node* belongs to. On the other hand, two *communities a* and *b* can share $s_{\mathrm{a,b}}$ *nodes*, which is defined as the overlap size between these *communities*. *Communities* also constitute a network at a higher level with the overlaps of proteins as their links or *degree*. The number of such links of a *community a* is called its *community degree d(a)*. The cumulative distribution functions denoted by $P(s_{com})$, $P(d_{\mathrm{com}})$, $P(s_{ov})$, and *P(m)* are calculated. The authors suggested a value of k between 4 and 6 [Palla et al., 2005] or selected the smallest value of k where no giant *community* appears.

4.2.2 The Network Clustering Coefficient

The *Network Clustering Coefficient* CC_n is introduced here to measure the linkage density of a given network. The CC_n describes the fraction of the number of *edges* to the full-connected subnetwork and is applied to any examined *community* to evaluate the connectedness.

The number of *edges* of a complete and full-connected network of V vertices is

$$\frac{V(V-1)}{2} \tag{4.1}$$

To ensure that the coefficient for a minimally connected graph is equal to zero, *V-1* is subtracted. Thus, the *Network Clustering Coefficient* is yielded by

$$CC_n = \frac{E-(V-1)}{\frac{V(V-1)}{2}-(V-1)} \tag{4.2}$$

with $0 \leq CC_n \leq 1$.

4.2.3 Biological Annotation of Communities

The annotation of *communities* to biological functions and/ or structures is manually done by evaluating the biomedical literature provided by NCBI PubMed or the diverse biomedical literature published by journals. Alternatively, GeneOntology [Ashburner et al., 2000] classification is applied to larger datasets.

4.3 Results

4.3.1 Biological Annotation and Local Structure of Communities

See supplementary.

Network Density of *Communities*

The CC_n is developed here to examine and compare network density. It is shown that full-connected complexes ($CC_n = 1$) represent mostly *structural* complexes. However, one full-connected complex, important for signal transduction, is found as well (see chapter 5).

Protein Interaction Predictions

The following interactions are lacking to complete the subgraphs to full-connected cliques and may thus serve here as defective cliques [Yu et al., 2006] and thus as predictions for putative interactions (table 4.1). First literature analysis showed that the predicted interactions are not reported yet.

Table 4.1: Protein Interaction Predictions.

Biological Function	predicted PPI	literature survey
The Exosome Complex	EXOSC5 – EXOSC4	not reported
COP9 Signalosome	TP53 – CUL5	not reported
NF-κB	IKBKG - RELA	not reported
TGF-β	TGF-β3 – TGF-β1	
Proteasome	PLK1 – SLC2A4	not reported
	PLK1 – PSMA2	
	PSMA3 – PSMA2	
TCP1 ring complex (TRiC)	CCT7 – CCT6A	not reported

4.3.2 Global Community Structure

It is shown that many *communities*, regarded as structural and/or functional *modules*, are connected to perform a higher level cellular organisation [Palla et al., 2005] that leads to a subsystems view. No interactions among the *communities* k11, k10 and k9 are observed here in the analysed human PPI network. The *communities* from k8 to k3 form networks that are represented as a graph $C=C(N^{com}, E^{com})$ and parameters are shown in table 4.2.

Table 4.2: Overview of the *community* networks at different *k*. N is the number of identified *communities*, N^{com} is the number of interacting *communities* and f describes the fraction of $N^{com}/\ N$ in per cent. The number of *edges* among the *communities* is denoted by E^{com}, CC_n is the *Network Clustering Coefficient,* $<d^{com}>$ the average *degree* and $<CC1(v)^{com}>$ the average Clustering Coefficient. *Comp.* is the number of connected components.

k-community	8	7	6	5	4	3	yeast k4*
N	14	16	41	97	235	400	82
N^{com}	8	11	34	86	210	364	-
f	57,14	68,75	82,93	88,66	89,36	91	-
E^{com}	17	14	58	272	796	850	-
CC_n	0,47	0,09	0,05	0,05	0,03	0,01	-
$<d^{com}>$	4,25	2,545	3,412	6,326	7,581	4,67	1,54
$<CC1(v)^{com}>$	0,708	0,615	0,548	0,624	0,61	0,538	0,17
comp.	1	2	5	3	3	3	-

* taken from Palla et al. and references therein; yeast network from the DIP core list (V=2609, E=6355)

Four parameters either monotonically increase(N, N^{com} and f) or monotonically decrease (CC_n). However, the other four parameters $<d^{com}>$, $<CC1(v)^{com}>$, E^{com} and the number of components, show non-monotonical progressions.

Regarding the average *clustering coefficient* of the *community* structure, the $<CC1(v)^{com}>$ is above 0,54 for any *k-core community*, compared to other available data (see table 4.3), denoting a highly interwired *community* structure for this human network.

Table 4.3: Comparison of N^{com}, $<d^{com}>$, $<CC1(v)^{com}>$ of other complex networks [Palla et al., 2005].

	N^{com}	$<d^{com}>$	$<CC1(v)^{com}>$	reference
co-authorship	2450	12,1	0,44	[Palla et al., 2005]
word assoc.	670	11,33	0,56	
yeast PPI k4	82	1,54	0,17	

Figure 4.1 shows the progression of $<CC1(v)^{com}>$ from k8 to k3. A small increase and a subsequent decrease is observed from k6 on. The number of components increases to k6 and stabilizes from k5 to k3 with 3 connected components (see table 4.2). This observation might reflect interesting biological properties and are further examined in chapter 5 in detail.

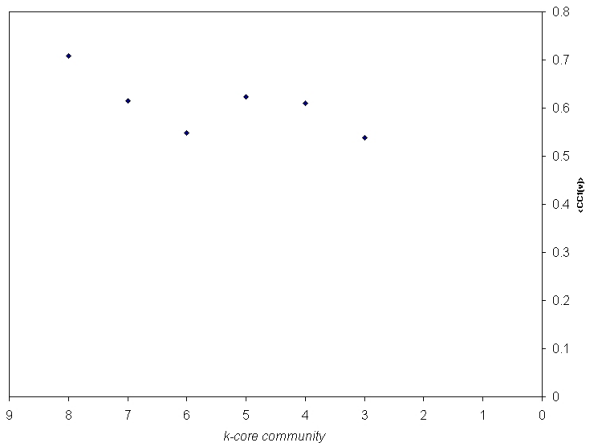


Figure 4.1: The average clustering coefficient of the *community* network at different k.

Community Sructure k8

Eight of the 14 *communities* are interlinked (see table 4.4 and figure 4.2). *Community* 8 is the most central comprising 19 proteins. This subnetwork has a nested structure with a k9 signaling core embedded (see chapter 5).

The structure reveals that c6, which is enriched in the STAT proteins 1/3/5a/5b (Signal Transducer and Activator of Transcription), links the other signaling *communities* to the Transcription Factors *communities* c12 and c13 (fig 4.2 and chapter 5).

Table 4.4: *Communities* at k8 with associated parameters and their biological activity. *Communities* 6 to 13 interact with each other and are annotated to signal transduction and transcription factors.

k	c	V	E	CC_{n}	Biological Activity	Reference
8	0	8	28	1	Actin-related protein complex ARP2/3	[Welch, 1999]
8	1	10	45	1	Histone Deacetylase Complexes (HDAC), Sin3A	[Minucci & Pelicci, 2006]
8	2	11	55	1	TBP and associated factors (TAFs)	[Burley & Roeder, 1998]
8	3	10	44	0,97	Exosome complex	[Houseley et al., 2006]
8	4	8	28	1	Sm-like proteins (LSM)	
8	5	10	44	0,97	COP9 Signalosome	[Wolf et al., 2003]
8	6	10	42	0,92	Signal Transduction (4 STATs)	
8	7	8	28	1	Signal Transduction	
8	8	19	113	0,61	Signal Transduction (with nested core k9)	
8	9	9	35	0,96	Signal Transduction	
8	10	8	28	1	Signal Transduction	
8	11	8	28	1	Signal Transduction	
8	12	10	43	0,94	Transcription Factors (3 Smads)	
8	13	8	28	1	Transcription Factors and Signal Transducers	

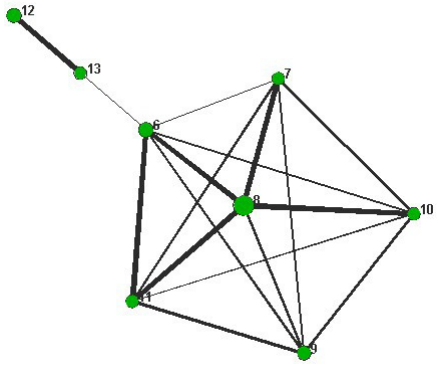


Figure 4.2: k8 *community* structure. Nodes represent *communities* and *edges* denote an overlap of proteins between *communities*. The size of the *nodes* and *edges* reflect number of *nodes* and number of overlapping proteins, respectively.

Community Sructure k7

A similar organisation as found for k8 is observed for k7. Eleven of 16 *communities* form a "two-component" network, 9 are in one main component. The *community* k8 c8 is further enriched with signaling proteins (and nested) to form one larger signaling component k7 c9 consisting of 62 proteins and 503 interactions. The transcription *modules* k8 c12 and c13 are fused in k7 c11 forming a larger transcription and translocator *module* with 22 proteins and 123 interactions (table 4.5 and figure 4.3).

The two largest *communities* c9 and c11 form a molecular bridge between signal transduction proteins in the cytoplasm and transcription factors in the nucleus, especially c11 is enriched with translocator proteins that redistribute from cytoplasm-to-nucleus upon activation (see chapter 5).

Table 4.5: *Communities* at k7 with associated parameters and their biological activity.

k	c	V	E	CC_{n}	Biological Activity	Reference
7	0	8	28	1	Actin-related protein complex ARP2/3	[Welch, 1999]
7	1	7	21	1	Mediator complex	[Belakavadi & Fondell, 2006]
7	2	10	45	1	Histone Deacetylase Complexes (HDAC), Sin3A	[Minucci & Pelicci, 2006]
7	3	7	21	1	Nucleotide excision repair NER incl. GTF2H1(TFIIH)	[Park & Choi, 2006]
7	4	10	41	0,89	Small nuclear ribonucleoproteins (snRNPs)	[Nilsen, 2003]
7	5	11	55	1	TAF	
7	6	11	51	0,91	Exosome	
7	7	9	34	0,93	LSM	
7	8	10	44	0,97	COP	
7	9	62	503	0,24	Signal Transduction	GO
7	10	7	21	1	n.d.	
7	11	22	123	0,49	Translocators (Smads, MAPK1, STATs)	[Xu, 2006]
7	12	8	27	0,95	NF-κB	GO
7	13	8	27	0,95	n.d.	
7	14	7	21	1	n.d.	
7	15	7	21	1	n.d.	

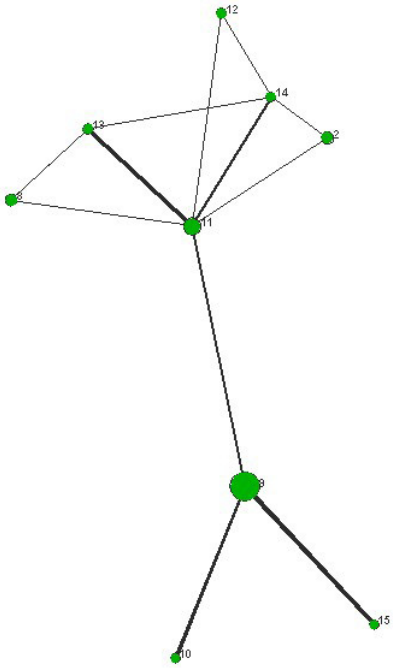


Figure 4.3: k7 *community* structure.

Community Sructure k6 and k5

Table 4.6 shows a subset of k6 and k5. In total 34 of 41 *communities* interact at k6 and 86 of 97 at k5. The largest *communities* at k6 are annotated to transcription (c24) and signal transduction (c12). At k5 one giant *community* appears where signaling and transcription fuses to one giant component (371 proteins). The theoretical properties of signaling and transcription *communities* are of special interest and are further analysed in chapter 5.

Table 4.6: Subsets of *communities* at k6 and k5 with associated parameters and their biological activity.

k	c	V	E	CC_{n}	Biological Activity	Reference
6		7			TGF-β	
6	24	48	305	0,24	43/48 Transcription DNA dependent	GO
6	12	107	947	0,15	95/103 signal transduction	GO
6		7			Apoptosis 7/7	GO
6	5	8	25	0,86	6/8 Proteasome Core Complex	31
6	20	6	15	1	B-cell antigen receptor complex	39
5	1	9	26	0,64	Replication Factor	33
5	12	6	14	0,9	TCP1 ring complex (TRiC)	34
5	13	5	10	1	CD3-TCR complex	
5	7	371	2730	0,03	215/350 signal transduction and transcription factors	GO
4	2	1098	6850	0,01	468/ 1024 signal transduction	GO

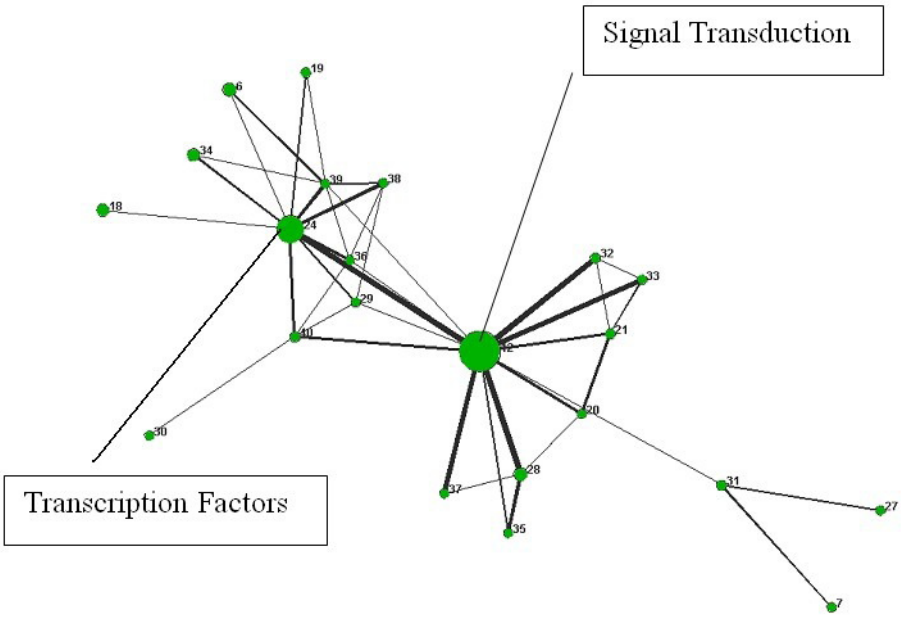


Figure 4.4: *Communities* at k6.

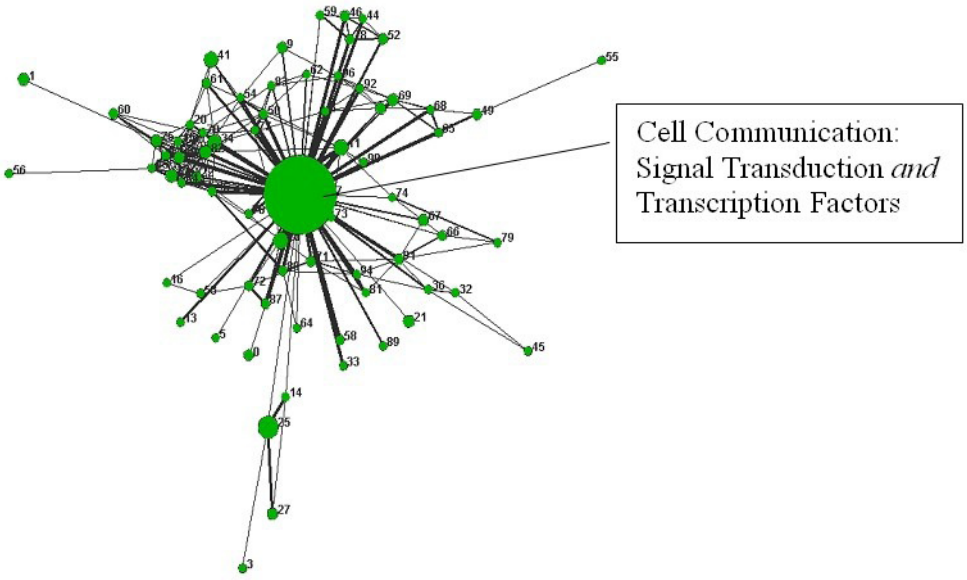


Figure 4.5: *Communities* at k5. The giant *community* in the center reflects a fusion of signal transduction proteins and transcription factors as a huge cellular communication module.

Regarding figure 4.4 and 4.5 the observation of figure 4.6 is now explained by the fusion of the largest *communities* for signal transduction and transcription factors.

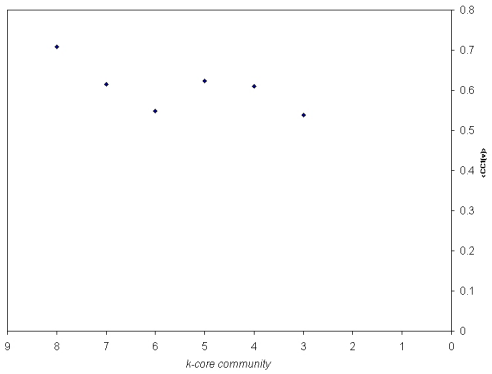


Figure 4.6: The average clustering coefficient of the *community* network at different k.

The four cumulative distributions for the *community* structure [Palla et al., 2005] are analysed and the statistical properties are compared and presented in the following.

Size Distribution

The cumulative distribution function of the *community* size at k7, k6 and k5 is shown in Figure 4.7. Palla et al. found a power-law dependence with an exponent of -1 and -1,6 for two non-biological complex networks and observed a lack of a specific scaling of their examined biological network (yeast), although a power-law like relationship is viewable when disregarding the tail with the four largest *communities*. The human network presented here shows also a power-law distribution up to the two largest *communities*, with a large exponent of -2,8, -2,7 and -2,3 respectively (see fig 4.7). However, the tail looks exponential.

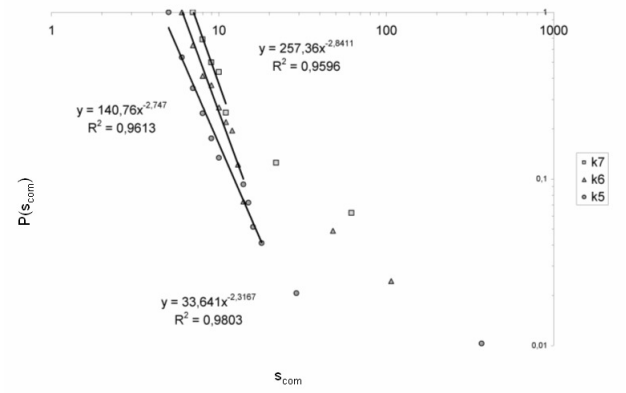


Figure 4.7: Cumulative size distribution at k5, k6 and k7.

Degree Distribution

Figure 4.8 shows the cumulative degree distribution at k7, k6 and k5 for the human network. At k7 and especially at k6 the specific scaling discovered by Palla et al. is also observed for the human network. The *community degree* has a unique distribution consisting of two parts, an exponential decay followed by a power-law tail. From k5 on one giant component emerges explaining the different distribution (see fig. 4.8).

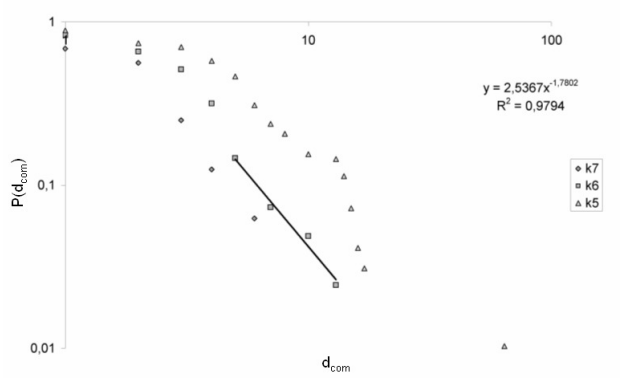


Figure 4.8: Cumulative *degree* distribution at k7, k6 and k5.

Overlap Size Distribution

Figure 4.9 shows the distribution of the overlap size. The largest overlap size is 17 at k4 and 9 at k6 (largest in yeast is 5 at k4 in Palla et al.). The distribution pattern for yeast and the two non-biological examples is close to a power-law [Palla et al., 2005]. However, the distribution of the overlap size for the human network looks rather exponential than power-law like, especially at k6 and k5.

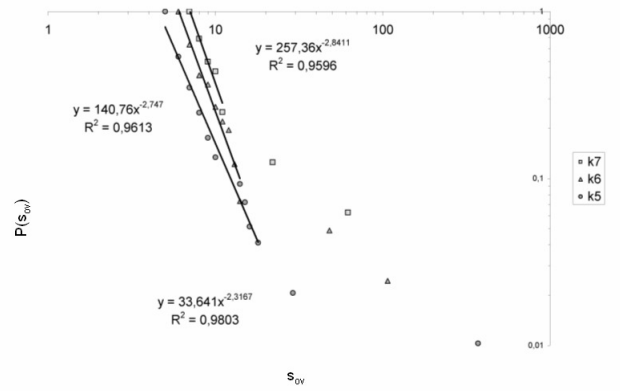


Figure 4.9: Cumulative distribution of the overlap size.

Membership Distribution

The highest membership number is 27 at k4 and 14 at k5 (yeast is 4)(fig. 4.10). A power-law distribution is shown for k8 to k4 with an exponent between -2,5 and -3,1 (table 4.7).

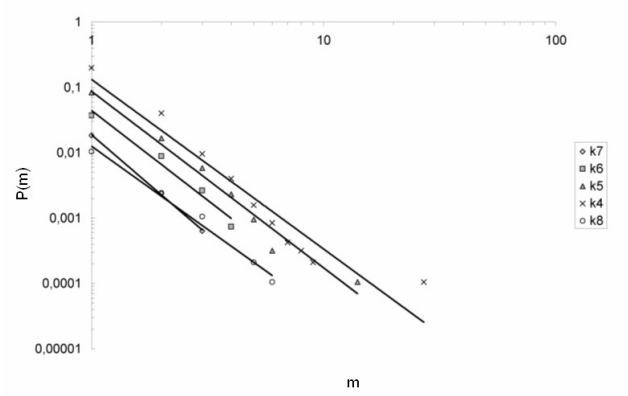


Figure 4.10: Membership distribution from k8 to k4.

Table 4.7: Exponents of the power-law membership distribution for the *community* structure.

k	8	7	6	5	4
exp.	-2,54	-3,05	-2,74	-2,7	-2,6

4.4 Discussion

The *CPM* method successfully finds and groups functionally relevant proteins representing regional network characteristics that mirror biological functions in the human PPI network. Known multiprotein complexes as well as functional *modules* are found. Complete coherent sub-processes are reconstituted by pure topological aspects, like the proteasome α subunit, the TFIID-complex and the intricate organisation of transcription initiation. Adamcsek [Adamcsek et al., 2006] and Zhang [Zhang et al., 2006] found as well that the *CPM* method successfully identifies functional *modules* in yeast PPI networks.
It is supposed here, and in well-agreement with the biological literature [Zhang et al., 2006], that most of the non-structural functions are organised less rigidly and are composed of more flexible subnetworks. Moreover many functions strongly overlap. Futschik et al. [Futschik et al., 2007] found as well functional *modules* using this method in a human PPI network. However, only a few large *modules* for transcription initiation and signal transduction are annotated automatically using Gene Ontology.
Lowering k integrates more *modules* and thus complete biological processes like transcription initiation or RNA manipulation can be uncovered at k5, as complex functions are build from smaller specialized modules. Also smaller *modules* at k4 are of high value and represent coherent functional complexes. This study

does not support that a specific k is of most value [Palla et al., 2005], this might be due to the rich *community* structure found in this human network. As structural *modules* are mainly represented by complete subgraphs found for k>7, the integration of smaller *modules* are found at lower k conditions. The largest found *communities* are annotated to intracellular communication (signaling and transcription factors) and dominate the whole structure from k5 on (see also chapter 5 for a detailed analysis of communication processes). Very interestingly, most of the *communities* are linked to the central communication *module* at k4. The various and numerous modules have to be organised and activated or repressed so that *modules* are able to *communicate* among each other.
Some preliminary predictions are drawn here from the perspective of defective-cliques, where *edges* that complete subgraphs to full-connected cliques, may function as predictions for yet unresolved protein interactions [Yu et al., 2006]. However, the subnetworks may be already well-extracted from literature such that no prediction is confirmed by a first literature survey here. On the other side, the subgraph may not comprise its activity as a complete clique.
Many biological functions are not annotated so far or less examined in the literature. Proteins associated with functionally classified *communities* may be interpreted as a prediction of their functions in the sense of *guilt-by-association*. It is shown that a method for adding proteins that enhance the subnetwork density is of value for new interpretational efforts as the *CPM* method might be too rigid to discover less examined or clustered processes. Moreover, embedded proteins may serve as functional predictions. Further interpretation and annotation especially for functional complexes of k6 and k5 is worthwhile.

It is observed here that regarding single *communities* as well as unions of the same give rise to numerous new insights into the organisation of functions. Different coherent biologically active functions are resolved here solely by topological examinations and thus the results are virtually biologically hypothesis-free. Additionally, the identified *communities* are important for further functional annotations and biological predictions.

Chapter 5

Topological Aspects of Signal Transduction

Abstract

Signal transduction networks enable the dynamic response to extracellular signals and provide high flexibility with only a limited set of signaling proteins. Understanding this complex organisation will give new insights into the molecular aspects of communication and into many human disorders that result from malfunction in signal transduction. The functional coherence of the largest found *communities* by the *CPM* method in the human PPI network is resolved here showing a distinct association to signal transducers and transcription factors as highly-connected subgraphs. A protein conjunction between these *communities* and compartments, representing biologically well-known shuttling proteins is topologically revealed. Moreover, an inner full-connected signaling core (and a highly interwired transcription core) surrounded by less interwired hierarchical layers of signaling (transcription) proteins is observed. The molecular nature of these proteins is largely associated to EGFR-signaling important for regulating proliferation, growth, survival and differentiation and is thus associated with cancer.

The presence of this highly interwired regions is also monitored by a high number of redundant *shortcuts* determined by the global communication kernel. Moreover the degree-degree dissortative correlation is very weakly pronounced opposed to other published biological examples. The nearly absence of this correlation is at least partly due to the observed cores consisting of many hub-proteins that form the two functional and compartmentalized dense subnetworks for cell communication. This top-hub integration points, which are recently also called *rich-clubs*, are revealed for the human network.

It is proposed here that signal transduction (and transcription) proteins, especially for EGFR downstream pathways, may be structurally organised at least partly as nested hierarchical shells that form a dense *rich-club* to accomplish a high diversity of signaling pathways and *cross-talk* with a restricted number of proteins. The presented topological features may also lead to new insights for cancer research as disturbed EGFR-downstream processes are highly central for cancer.

5.1 Introduction

Signal transduction is an important biological process by which cells transduce external stimuli into a cellular response. Environmental signals can specifically be detected by cell membrane receptors which activate a sequence of biochemical intracellular reactions that transfer the signal into the nucleus where gene expression is activated [Krauss, 2003]. Signaling pathways are complex in nature and conducted by a variety of protein protein interactions forming complex signaling networks. Although signaling is a highly dynamical process, the underlying static structure defines the pathway space and constraints. This chapter focuses on the largest found *communities* depicted by the method of *k-core communities* [Palla et al., 2005] in chapter 4. The biological annotation shows here a profound functional association to signal transducers and transcription factors. To get more insight into intracellular communication aspects, the *regional* properties of the found *communities* are analysed and interpreted. Further emphasis is made on a comparison of the subgraph densities, clustering properties and average degrees at different k. For finding other potential structural features for intracellular communication, the *regional* findings are opposed to *global* network examinations. Kim et al. [Kim et al., 2004] proposed that the *skeleton*, a minimal spanning tree (MST) based on *edge betweenness*, and added shortcuts, organise a complex network and the *shortcuts* are responsible for the *clustering* properties of a network. Interestingly, the hitherto published *skeletons* are all *scale-free*. The authors classified the *shortcut* length distribution, the number of pairwise *node* distance on the *skeleton*, into a longer-loop dominant structure (biological networks) and a monotonically decreasing structure (Internet). Both features are examined and confirmed here for the human PPI network. The functional annotation and interpretation of proteins with a high number of *shortcuts* shows as well a high content of signal transducers and transcription factors. This *global* finding is further examined by *local* observations.
Maslov and Sneppen [Maslov & Sneppen, 2002] examined the degree-degree correlation, the relationship of the *node degree* to the average *degree* of direct neighbours. The *degree* correlation is classified into three types (see methods) and it is found that biological networks typically exhibit the dissortative (repulsive) property, where high- *degree nodes* tend to interact with low-degree neighbours. This property is analysed and confirmed here for the human network. Some proteins that *smooth* or *deviate* the dissortative property are biologically annotated to signal transducers and transcription factors that are also revealed with the *CPM* method as central cores as well to proteins having high-shortcuts.
[Luo et al., 2007] examined *rich-clubs*, dense subnetworks of hub-proteins as central traffic hubs for the fast integration and transmission of information through the network. Although the notion and the measuring of *rich-clubs* is a current debate [Luo et al., 2007], [Colliza et al., 2006], a dense *rich-club* for the human network is found and shown to overlap with the signaling and transcription *community* cores, the high-redundant *shortcut* proteins on the *skeleton* and the hubs perturbing the dissortative correlation. This chapter is structured as

followed. The methods section describes the notion of a spanning tree and how the *skeleton* is calculated as well as the classification of the *degree*-correlation. The results section presents different overall statistical and network parameters and the functional annotation for the two *communities* under varying k. Overlapping proteins are interpreted as an own biological active conjunction or translocation *module* between the two *communities* and compartments. A nested structure signaling mechanisms is observed and the molecular basis of the inner most proteins is resolved. Subsequently, the scale-free property and the linear dependence of the *skeleton* compared to the original network is examined to get more insight into communication processes from a *global* view. Proteins having high numbers, as well as low numbers of *shortcuts* on the original network, are examined. Globally determined proteins with high redundant *edges* or shortcuts (low information load globally) are compared to the regionally established proteins of the nested shells (highly interwired regional layers) based on the *community*-method. The *degree* correlation is presented and its distribution is discussed. Finally, the presence of a *rich-club* for communication is shown. The last section discusses the findings and proposes further directions for the understanding of cell communication processes.

5.2 Methods

The network density method CC_n and the *k- core communities* described in chapter 4 are applied, methods for network measurements are presented in chapter 2.

5.2.1 The skeleton or communication kernel

The *skeleton* or communication kernel is a special type of a minimal spanning tree (MST). An MST has *V-1 edges* and no shortcuts. To delete all shortcuts, a weighting of *edges* was chosen due to the *betweenness centrality* as a sum of the *betweenness centrality* of incident *vertices*.

For the calculation of the MST, *edges* are weighted by

$$w(e_{u,v}) = BC(u) + BC(v) \tag{5.1}$$

The MST is calculated by maximizing the total *weight* of the *edges*

$$w(e_{u,v}) \leq w(e_{u+1,v+1}) \tag{5.2}$$

with u and v being incident vertices.
For algorithms, see Sedgewick et al. [Sedgewick & Wyk, 2002]

5.2.2 The degree-degree correlation

Scale-free networks are grouped into three classes according to *node degree* and *degree* of its direct neighbours: assortative (large *degree nodes* tend to connect to large *degree neighbours*, e.g. social networks), dissortative (large *degree* connects to small *degree* neighbours, e.g. internet and biological networks) and neutral (no correlation, e.g. in silico-networks) [Goh et al., 2003].

5.3 Results

5.3.1 Communities of Cell Communication

Community Properties at different k

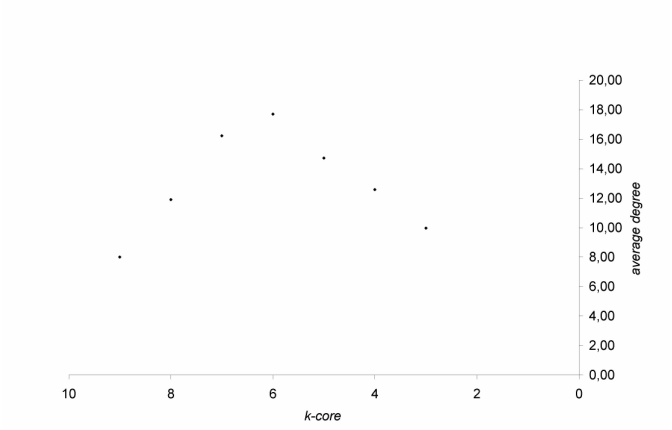
During the analysis and annotation of the *communities* in chapter 4, it is noticed that a few *communities* are part of larger ones when regarding a lower k-level. This observation is systematically examined in more detail in the following.

The comparison of *community* parameters reveals that the average *clustering coefficient* $<CC1(v)>$ decreases to k6a, increases slightly at k5 and further decreases to k3 (see fig. 5.1b). In turn, the average *degree* $<d>$ increases to k6a and further decreases to k3 (see fig. 5.1b). In the opposite, the network density CC_n decreases and the diameter increases from k9 to k3 (table 5.1).

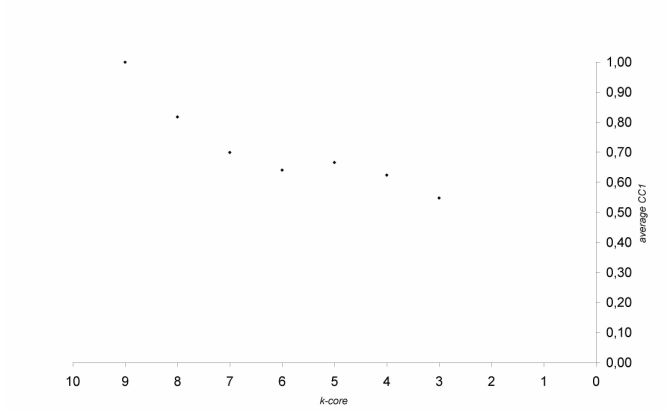
Table 5.1: Basic theoretical properties for the largest *communities* at different k. Note that the $<CC1(v)>$ is still 0,55 (55%) at k3. The *diameter* of the whole giant component is 12.

k-core	*V*	*E*	CC_n	$<d>$	$<CC1(v)>$	*diameter*
9	9	36	1	8	1	1
8	19	113	0,62	11,89	0,82	2
7a	62	503	0,24	16,23	0,70	3
6a	107	947	0,15	17,70	0,64	3
5	371	2730	0,03	14,72	0,67	5
4	1098	6850	0,01	12,58	0,62	7
3	3634	18115	0,00	9,97	0,55	9

Up to k6 more *edges* related to the number of *vertices* emerge and subsequently the relation is inverted, more *vertices* than *edges* arise (fig. 5.1), reflecting that the *communities* get more dense up to k6.



(a) <d> in dependence of the largest k-core *communities*



(b) <CC1(v)> in dependence of the largest k-core *communities*

Figure 5.1: Dependencies of the largest *k-core communities*

The distributions suggest that the *community* properties change between k6 and k5. To examine the difference in network properties between k6 and k5, the presented *communities* are functionally annotated.

Functional annotation of large *communities*

Biological annotation shows a *distinct* functional association to signal transduction from k9 to k6 (table 5.2). Interestingly, at k5 signal transduction proteins as well as proteins for transcription overlap and integrate into one *community*. The proteins of k6a are a proper subset of the proteins of k5.

Table 5.2: Functional Annotation of 7 large *communities* at different *k*. Classification is done with GeneOntology. Note that some numbers of proteins vary between tables, because not any protein is classified by GeneOntology. k5 is also annotated to cell proliferation (51 proteins) and basic nucleic acid metabolism (141 proteins).

k-core	V	E	signal transduction	transcription
9	9	36	9/9	
8	19	113	19/19	
7a	62	503	58/ 60	
6a	107	947	95/103	
5	**371**	**2730**	**215/350**	**58**
4	1098	6850	468/ 1024	n.d.
3	3634	18115	n.d.	n.d.

Additional *communities* for transcription factors are annotated and the network properties are examined (table 5.3) to get more insight into *communities* for transcription factors.

Table 5.3: Network properties of *communities* for transcription factors.

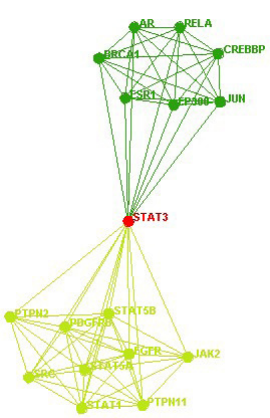
k-core	*V*	*E*	transcription	CC_n	*<d>*	*<CC1(v)>*	*diameter*
8b	10	43	10/10	0,94	8,5	0,95	2
8c	8	28	8/8	1	7	1	1
7b	22	123	21/22	0,49	11,18	0,79	2
6b	48	305	43/ 48	0,24	12,71	0,7	3
5	**371**	**2730**	**58/ 350**	**0,03**	**14,72**	**0,67**	**5**

The next section describes the molecular basis of the overlapping proteins within a k-core *community* level and the overlaps between the different k-core levels.

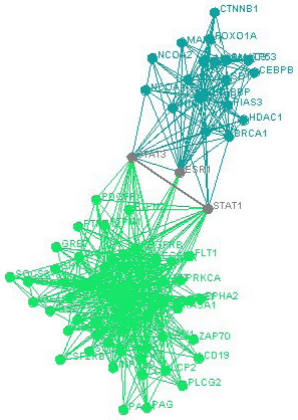
Overlapping proteins between *communities* of the same k-core level

The signaling pathways of the epidermal growth factor receptor EGFR are important to regulate proliferation, growth, survival and differentiation in mammalian cells[Oda et al., 2005]. Signaling and Transcription Activation are spatially separated in the cytoplasm and the nucleus. These two compartments are found here as independent *communities*. Nevertheless the signal has to be transduced into the nucleus by gated transport through the NPC (see chapter 3). At k8, two *communities* are linked via STAT3 (fig. 5.2a). STAT proteins have evolved specifically for the task to provide a fast and reliable mechanism to transduce a signal into the nucleus, and moreover have the ability to activate transcription [Reich & Liu, 2006]. *Community* 6 of k8 is enriched with

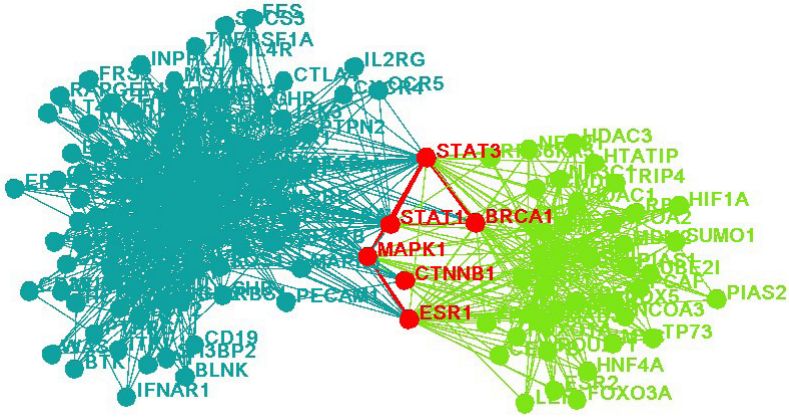
shuttling proteins between the cytoplasm and the nucleus and thus the pure topology mirrors well-known biological functions (fig. 5.2a).



(a) k8 c6 (below) and c13 (upper) overlap



(b) k=7 comm. 9 and 11



(c) k=6 *communities* 12 and 24

Figure 5.2: shows the molecular bridges revealed by the CPM method. a) STAT3 as overlap and (b) STAT1 and the estrogen receptor (ESR1) are included (c) shows β-catenin (CTNNB1), the MAP-kinase (MAPK1) and the breast cancer 1 protein (BRCA1).

The two largest *communities* c9 and c11 at k7 form a molecular bridge between signal transduction proteins in the cytoplasm and transcription factors in the nucleus, the overlap is enriched with translocator proteins that redistribute from cytoplasm-to-nucleus upon activation (STAT1, STAT3 and ESR) (fig. 5.2b).

At k6 even six proteins form the cellular link connecting the two spatially and functionally separated *modules* (fig. 5.2c).

STAT1/ 3/ 5A/ 5B

In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor associated kinases, and then form homo- or heterodimers that translocate to the cell nucleus where they act as transcription activators (NCBI Gene).

MAPK1

MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development. The activation of MAPK1 requires its phosphorylation by upstream kinases. Upon activation, this kinase *translocates to the nucleus* of the stimulated cells, where it phosphorylates nuclear targets (NCBI Gene).

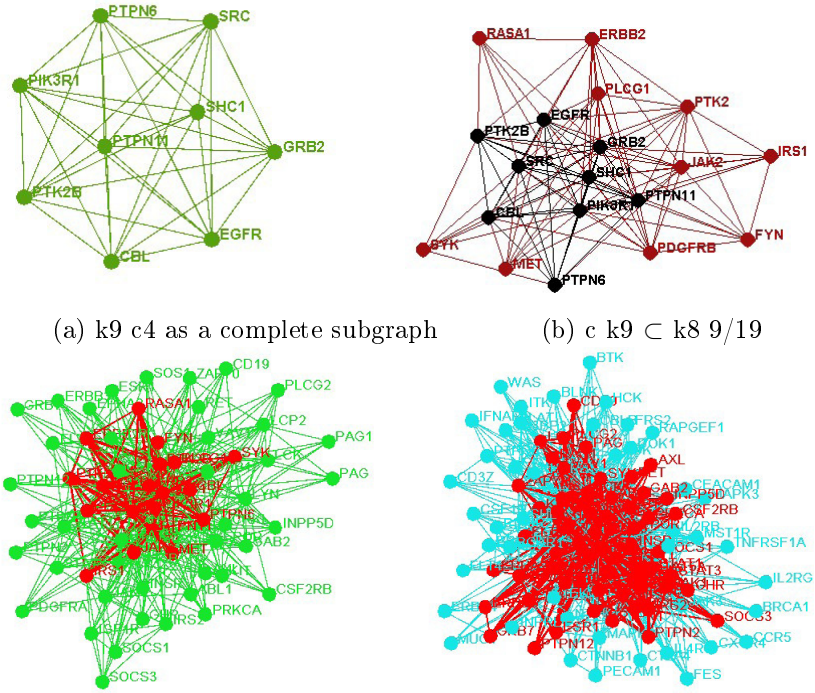
The subcellularly devided signal transduction and transcription is found by the *CPM* method by pure topological aspects. The overlap of proteins at *different k-core* levels are examined in the following section.

Overlaps between k-core levels

It is revealed that the largest *communities* are nested, smaller *communities* are part of larger ones (see table 5.4) of a lower k-level. Although the principle of the *CPM* method calculates complete subgraphs that overlap (chapter 4) and thus share *nodes*, only the largest *communities* of the respective k-level reveal such a nested structure.

Table 5.4: *Communities* with larger k are proper subsets of lower k *communities*.

subset	*number*
k6a⊂k5	107/371
k6b⊂k5	48/371



(a) k9 c4 as a complete subgraph

(b) c k9 ⊂ k8 9/19

(c) c19 k8 ⊂ k7a c9 19/62

(d) c9 k7a ⊂ k6a c12 62/107

Figure 5.3: Nested *communities* revealed for larger *modules*.

Transcription Factors

Two large *communities* for transcription are found (fig. 5.4 and 5.5) which are proper subsets with a fusion of two higher connected *communities* forming an inner transcription core. Annotation is done with GeneOntology.

Table 5.5: Transcription *communities* are proper subsets of larger *modules*.

subset	*number*
k8c⊂k7b	10/22
k6b⊂k5	48/371

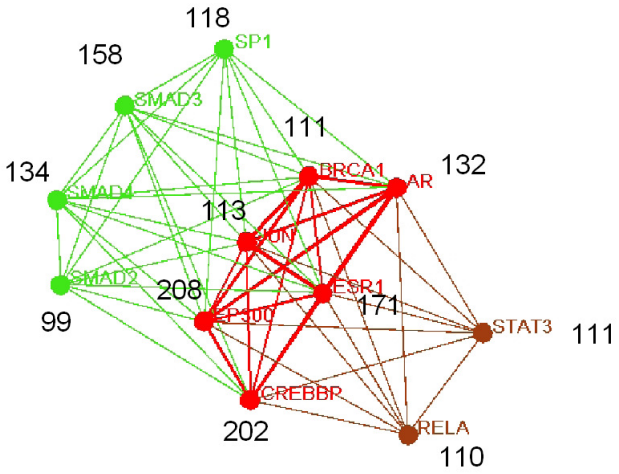
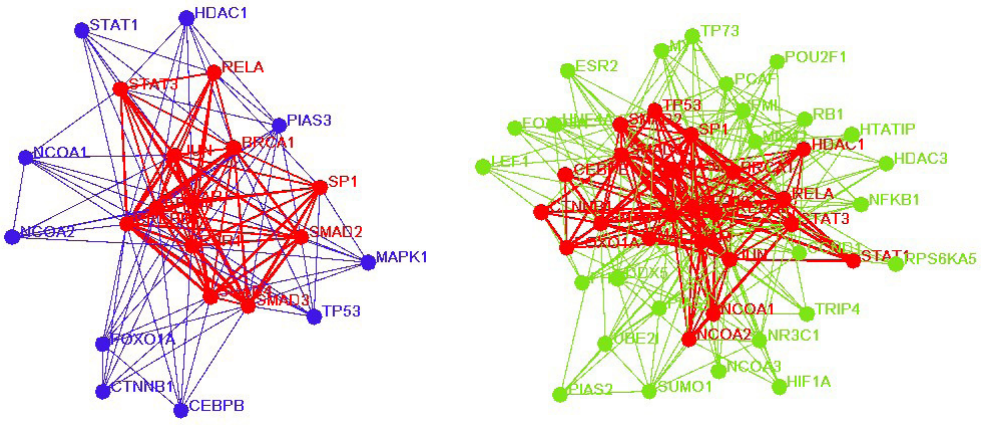


Figure 5.4: k8b c12 and k8c c13 transcription *modules* with degrees on original network showing a top-hub transcription module

Twelve transcription and top-hub proteins (average *degree* is 139) form a tightly interlinked network (see fig. 5.3 and 5.4).



(a) c12 and c13 k8b ⊂ k7b c11 (12/22) (b) c11 k7b ⊂ k6b c24 (22/48)

Figure 5.5: c12 and c13 k8b ⊂ k7b c11 (12/22) and c11 k7b ⊂ k6b c24 (22/48) transcription *modules* with degrees on original networks.

Figure 5.6 shows a schematic diagram of the communication hierarchies.

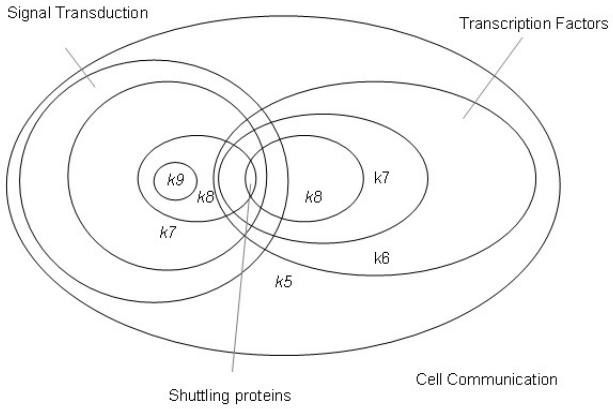


Figure 5.6: A schematic diagram of the cellular communication shells.

It is observed, that the four nested signaling *communities* are hierarchically organised, as the clustering properties decrease with the *degree* (see fig. 5.7).

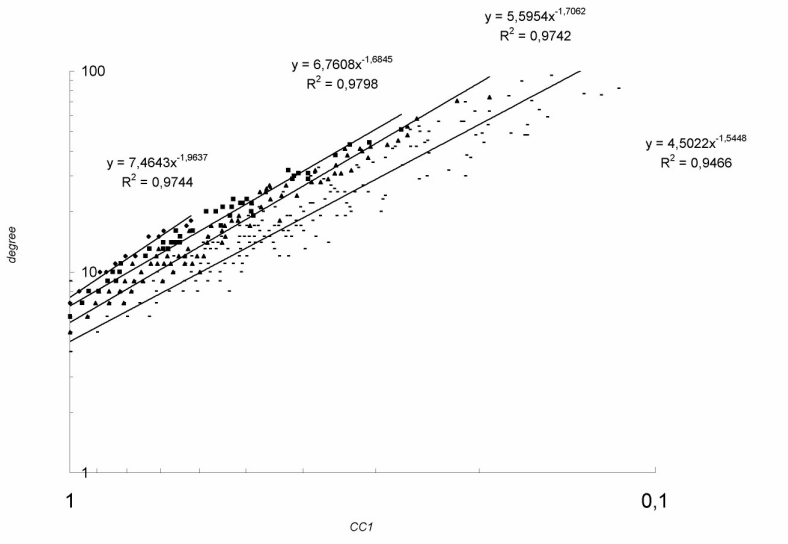


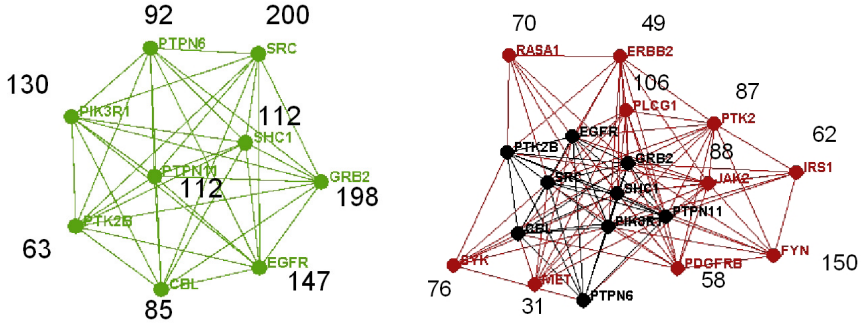
Figure 5.7: Function of *degree* with clustering properties of the signaling shells k8 to k5. Note that k9 is full-connected and thus the CC1(v) vanishes.

Molecular Basis of the inner signaling core (Epidermal growth factor receptor signaling)

The proteins of the inner core are associated to signaling pathways of the epidermal growth factor receptor EGFR which is important to regulate proliferation, growth, survival and differentiation in mammalian cells [Oda et al., 2005]. All proteins of the inner core (see table 5.6) are also enriched in SH2 and SH3 domains which are well-known for fast, reliable and effective binding of signaling proteins [Alberts, 2002].

Table 5.6: Molecular function of the full-connected signaling core.

protein	molecular function
GRB2	signaling adapter
CBL	signaling adapter
SHC1	signaling adapter
PIK3R1	signaling adapter
PTPN6	tyrosine phosphatase
PTPN11	tyrosine phosphatase
SRC	tyrosine kinase (membrane)
PTK2B	tyrosine kinase (cytoplasmic)
EGFR	receptor/ ser/thr kinase



(a) the central signaling core k9 c4 (b) the second signaling shell at k8

Figure 5.8: Degrees of the original network.

The two tyrosine specific protein phosphatases (PTPase) *PTPN6* and *11* (also known as SHP1 and 2) are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation (NCBI Gene). *GRB2* is an **adapter protein** that provides a critical link between cell

surface growth factor receptors and the Ras signaling pathway. It associates with activated Tyr-phosphorylated *EGF* receptors and *PDGF* receptors via its SH2 domain. *CBL* is also a well-known **adapter protein** with SH2 and SH3-domains and plays also a key role in immune response [Machida & Mayer, 2005]. In response to a variety of growth factors, *SHC1* binds to phosphorylated Trk receptors through their phosphotyrosine binding (PID) and/or SH2 domains. The PID and SH2 domains bind to specific phosphorylated tyrosine residues. These phosphotyrosines act as docking site for GRB2 and are thus involved in Ras activation. SHC1 is phosphorylated by activated epidermal growth factor receptor and acts as a **signaling adapter** that couples activated growth factor receptors to signaling pathways. Defects in *PIK3R1* (Phosphatidylinositol 3-kinase regulatory subunit α) are a cause of severe insulin resistance. It binds to activated (phosphorylated) protein-Tyr kinases, through its SH2 domain, and acts as an **adapter**, mediating the association of the p110 catalytic unit to the plasma membrane. It is necessary for the insulin-stimulated increase in glucose uptake and glycogen synthesis in insulin-sensitive tissues (NCBI Gene). Wu et al. [Wu et al., 2001] concluded that SHP-2 (PTPN11) is required for **mediating** PI3K/Akt (PIK3R1) activation.
PTK2B (also known as focal adhesion kinase FAK2 or PYK2) has been shown to bind to the SH2 domain of GRB2 and activates the MAPK signaling pathway. Wiiger et al. [Wiiger & Prydz, 2004] proposed that EGFR, PYK2 (PTK2B), Yes, and SHP-2 (PTPN11) are involved in transduction of the tissue factor/ coagulation protease factor TF/FVIIa signal possibly via transactivation of the EGF receptor. EGFR and the non-receptor protein tyrosine kinases SRC and Pyk2 (PTK2B) have been shown to be implicated in linking a variety of G-protein-coupled receptors (GPCR) to the MAP kinase signaling cascade [Andreev et al., 2001].
Modeling and understanding the diverse molecular reaction events solely for the EGFR protein is a current research topic [Blinov et al., 2006]. The number of possible phosphoforms of EGFR is 29 where 9 is the number of amino acids that are subject to phosphorylation and dephosphorylation [Jorissen et al., 2003].

To examine if these shells might be global communication channels, the *skeleton* is calculated.

5.3.2 The skeleton or communication kernel

To understand network function and the observed nested *community* structures for cellular communication further, the *skeleton* is calculated. *Edges* that are most important for transferring signals are calculated by reducing redundancies and keeping *edges* with highest information load, *edges* which are frequently transferred for geodesic paths.

Correlation of *degree* on *skeleton* versus original network

A linear dependence for *degree* of the original network *d(v)* and *degree* on *skeleton* $d(v_s)$ is observed (table 5.7) for the human network by applying the *Pearsons Correlation Coefficient*. This correlation is also discovered for various scale-free networks from sociology, technology and biology (yeast) by Kim et al. The correlation denotes that the *skeleton* is a network with lower complexity but comparable node *degrees*.

Table 5.7: The Pearson Correlation Coefficient for *degree* of MST and *degree* on original network.

Network	*Pearsons Correlation*
human (this study)	0,784
yeast [Kim et al., 2004]	0,814

The *degree* distribution of the human *skeleton*

The calculated human *skeleton* shows a scale-free property (fig. 5.9) although less pronounced opposed to the original network. Goh et al. discovered this feature also for other scale-free networks namely the World Wide Web, the *Escherichia coli* metabolic network, and a small *Homo sapiens* PPI network. A very low probability for *nodes* to have a *degree* above 40 is observable ($P(d) = 10^{-4}$).

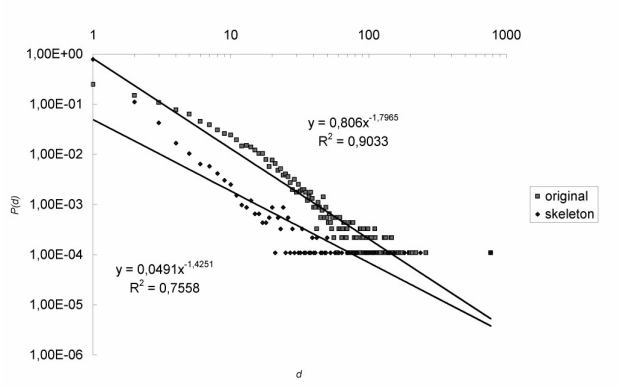


Figure 5.9: Degree and *P(v)* of the original network and skeleton.

This observation further supports the hypothesis of Kim et al. that a complex scale-free network consists of a scale-free tree and shortcuts added on it.
To examine the role of shortcuts, the authors examined the shortcut lengths,

which is the minimum number of steps between two given vertices on the *skeleton*, and observed two types of *shortcut* length distributions: the type I longer-loop dominant structure (coauthorship network, yeast PIN) and the type II structure, where the number of shortcuts decrease monotonically as the length increases (Internet), indicating a tree-like structure. Fig. 5.10 shows the distribution of the *shortcut* lengths for the presented human network.

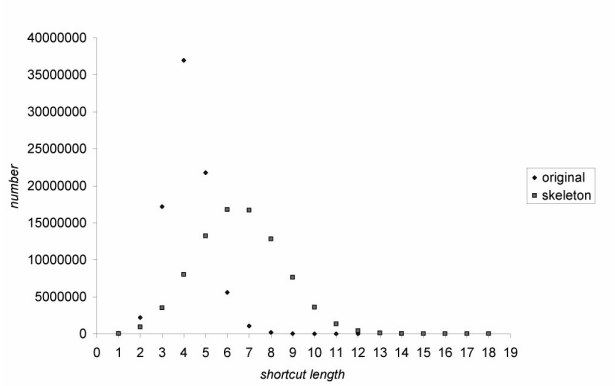


Figure 5.10: Shortcut length distribution of original network and skeleton.

The human *skeleton* and the original network belong to the type I longer-loop structure, which is similar to a Gaussian distribution and in accordance with Kim et al. Naturally, the average and highest distance is higher on the *skeleton* (no shortcuts). Compared to the original network, any pair is on average reachable in only six to seven hops on the *skeleton* (table 5.8) (increase of one third, also for the highest distance).

Table 5.8: Average and highest distance of shortcut length on original network and *skeleton* among all 85.036.062 pairs of the 9222 proteins.

	average distance
original	4,178
skeleton	6,541

The small average distance might be due to the scale-free structure of the skeleton.

To study the role of the top-hubs (and its associations) in maintaining the scale-free structure, the protein degrees of the two networks are compared.

Degree redundancy on *skeleton* compared to original network

The top hub protein SLC2A4 shows no change in *degree* k on MST compared to the original network (table 5.9) demonstrating the importance of top-hubs for network integrity. However, the other top-hubs do not retain their *degree* on the *skeleton*. Twelve of the thirteen top-hubs (degree ≥ 150) result in a *degree* higher or equal than 50 on the skeleton, except FYN (150 down to 27). Regarding the MST, these proteins are most important for maintaining the scale-free property because top-hubs remain top-hubs on the *skeleton*.

Table 5.9: Top-hub proteins with *degree* on original network and skeleton.

protein	original	skeleton
SLC2A4	770	770
TP53	261	237
YWHAG	221	183
EP300	208	81
CREBBP	202	153
SRC	200	141
GRB2	198	108
PRKCA	175	137
ESR1	171	70
MAPK1	166	76
SMAD3	158	72
CSNK2A1	151	50
FYN	150	27

Referring to the proteins having the highest absolute decrease of *degree* comparing the original network to the *skeleton*, proteins for signal transduction and transcription activation are found. The presented proteins in table 5.9 are part of highly-interwired regions having high redundancy by means of the number of *shortcuts*.
The comparison of high-shortcut proteins with nested shells proteins shows a strong overlap. Interestingly, all 9 proteins of the inner signaling shell are among the top 80 (6 among the top 32) and all 19 of the k8 shell are among the top 300 (16 among top 100), representing important signaling mediators. Some top-hubs for transcription and from the transcription core are present among the 20 highest *shortcut* proteins (table 5.10 and fig. 5.11).

Table 5.10: Top 20 proteins sorted by absolute decrease in *degree*. Five entries are part of the k9 signaling adapter complex (bold).

k	main	functional significance	*CC1(v)*	k_s	k_s/k	k-k_s
208	EP300	transcription core	0,0311	81	0,3894	127
150	FYN	signaling shell	0,0524	27	0,1800	123
122	**SHC1**	signaling core	0,1053	7	0,0574	115
112	**PTPN11**	signaling core	0,0988	2	0,0179	110
116	MAPK3	translocation	0,0337	9	0,0776	107
147	**EGFR**	signaling core	0,0494	45	0,3061	102
130	**PIK3R1**	signaling core	0,0756	28	0,2154	102
171	ESR1	translocation	0,0477	70	0,4094	101
151	CSNK2A1	signaling	0,0193	50	0,3311	101
132	AR	transcription core	0,0449	33	0,2500	99
106	PLCG1	signaling shell	0,0929	7	0,0660	99
133	RB1	transcription	0,0403	35	0,2632	98
134	SMAD4	transcription core	0,0347	38	0,2836	96
113	JUN	transcription core	0,0681	18	0,1593	95
99	LCK	signaling shell	0,0919	4	0,0404	95
99	SMAD2	transcription core	0,0583	8	0,0808	91
198	**GRB2**	signaling core	0,0449	108	0,5455	90
130	HDAC1	transcription shell	0,0515	40	0,3077	90
166	MAPK1	translocation	0,0334	77	0,4639	89
111	AKT1	signaling shell	0,0333	22	0,1982	89

From a biological perspective, this result is expectable as signaling proteins (and to some extent transcription factors) are highly interlinked. However, the result demonstrates that important cellular signaling proteins are identifiable by their topology by means of high shortcuts and their presence in the largest *communities*. Moreover, the *skeleton* based on the betweenness centrality is a feasible method for detecting biological signatures for communication processes.

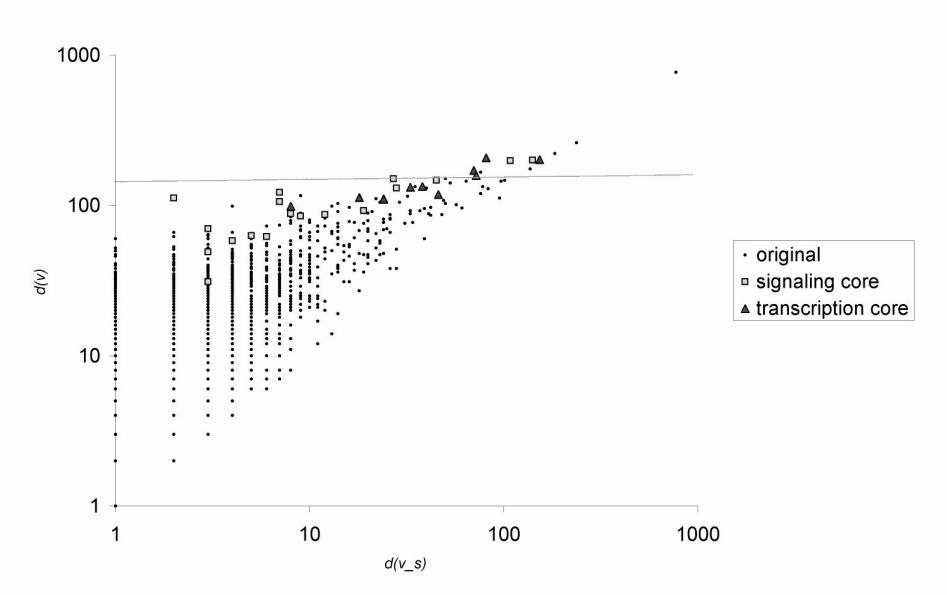


Figure 5.11: Reduction of k on spanning tree. With *d(v)* the *degree* on original and $d(v_s)$ on skeleton. The horizontal line denotes the *degree* above or equal 150 (note that there is no defined cutoff for hubs or top-hubs). The squares (core) denote the 19 proteins of the central and second signaling shell revealed by the *CPM*-method (*communities*).

In this section it is observed that the signaling and transcription cores are composed of hub proteins. The next section examines the impact of this finding for the global distribution of the *degree* correlation.

5.3.3 The degree-degree correlation

The degree-degree correlation is dissortative in the presented network (table 5.11 and fig. 5.12). Here, nine of the hub (*degree* > 84) signaling proteins work together to form a tightly interlinked integration point enriched with SH2 and SH3 domains. Moreover twelve transcription top-hubs form a second dense (although not full-connected) subnetwork (*degree* > 98). The figure 5.12 denotes that the 21 proteins from these inner cores affect slightly the distribution to a less pronounced dissortative property, towards a more neutral distribution. On contrary, the *degree* distribution of the *skeleton* is much more distinct, although highly scattered (fig. 5.12).

Table 5.11: Power-law exponent of the neighbourhood connectivity distribution (taken from Schueler [Schüler, 2006] and references therein).

	HIM	*FIM*	*YIM*	Stelzl et al.	Ito et al.
γ	$\sim$ -0.15	$\sim$ -0.06	$\sim$ -0.05	$\sim$ -0.32	$\sim$ -0.59

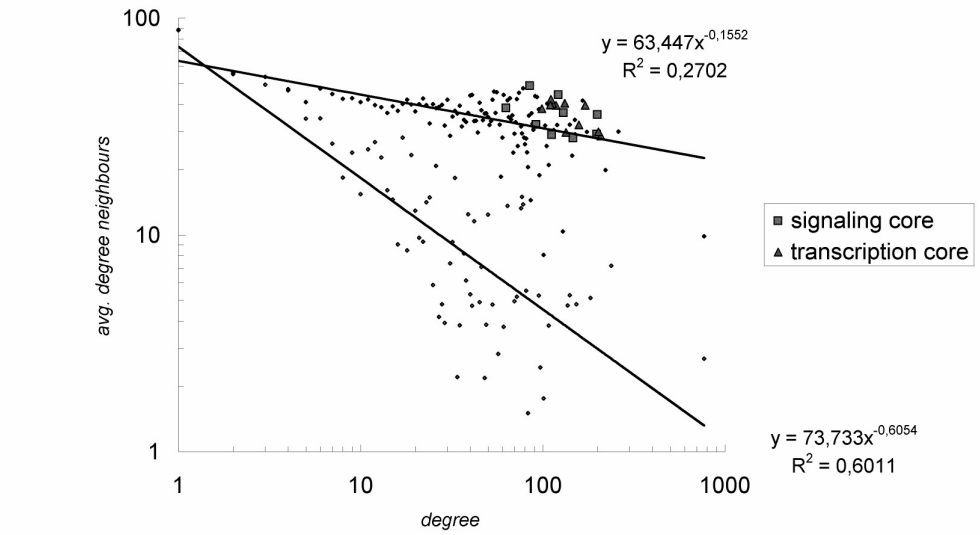


Figure 5.12: Degree-degree correlation of original network compared to the *skeleton* (lower line).

The coherence of the top-hubs (hubs) with *degree* $\geq$ 150 (at least >100), which is also referred to as a *rich-club*, is analysed in the next section. Moreover, as the *CPM* method might be too rigid, a connected component with the highest *degree* is calculated and analysed.

5.3.4 The *rich-club* phenomenon

The *rich-club* phenomenon describes the notion that highly connected *nodes* form a densely interlinked subnetwork dominating largely the structure and hence function of the network. The *rich-club* property was shown for a scientific collaboration network and shown to be absent in a yeast protein interaction

network[Colliza et al., 2006]. Here, the presence of the *rich-club* ordering is not examined by using the *rich-club coefficient* [Colliza et al., 2006], but by visualizing the proteins with *degree* larger than 100 (on original network), 42 of 44 proteins form a complex web of top-hubs (fig. 5.13). The TNF receptor-associated factor 2 TRAF2 and DIPA (also: CCDC85B) are not interwired. It is possible that human networks exhibit a *rich-club* ordering only for signaling proteins which would be in accordance with cellular communication. The observed dense region is split into two cores, connected by a third group that is associated to both and mediate between the two compartments and functions. It is functionally and spatially subdivided into signaling, translocation and transcription factors- mirroring the findings with the *CPM* method very well.

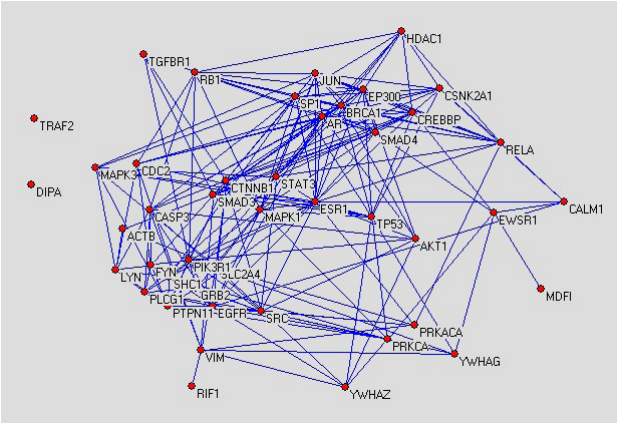


Figure 5.13: 44 proteins with *degree* above or equal 100 on original network forming a *rich-club*. It is denotable that at least some of these important proteins are associated with cancer (TP53, BRCA1, EGFR and MAPK1).

The 13 top-hubs (table 5.10) with a *degree* higher than 150 are still interconnected and form an own subnetwork (fig. 5.14). Only FYN and SLC2A4 have a *degree* of 1 in this top-hub network (fig. 5.14).

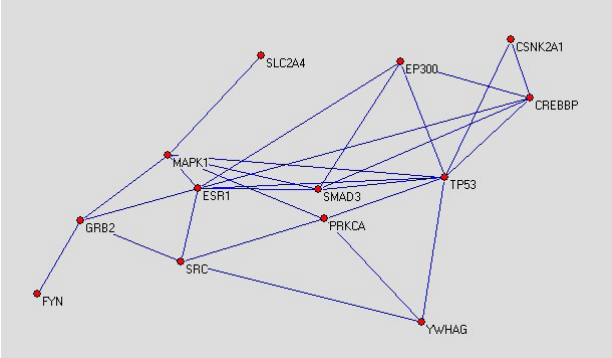


Figure 5.14: 13 proteins with *degree* higher than 150 on original network forming a connected backbone.

5.4 Discussion and Outlook

Intracellular signaling networks control a wide variety of cellular aspects, and are arranged through the diverse interactions of proteins with each other, with nucleic acids, with lipids or small molecules. The dynamic pathways are constricted to an intrinsic web of protein interactions as a structural backbone.

5.4.1 Communities

It is proposed here that signaling proteins comprise a specific regional and global signature as they are organised partly as nested core-periphery shells. This idea is further supported by the observation that inner-core proteins have also the highest global *edge* redundancy. This observation is in very good agreement with biology and expectable as signaling proteins have to be complexly interwired to exchange information fast and reliable. This hypothesis is also in well-agreement with the notion of *cross-talk*, where signaling pathways are complexly interwired and interchange protein usage. By utilizing proteins from higher connected and more central shells, information can be simply interchanged and directed to other pathways and vice versa. Bray [Bray, 1998] points out that clusters in permanent association as signaling complexes are well suited to perform a rapid, efficient and noise-free way for transducing signals. The complex could work repeatedly, processing signals rapidly and accurately without further need for diffusion. Modifications to the assembly reacting to the current state of the cell could be made without requiring the complex to be disassembled. Proteins like GRB2 have evolved solely for the purpose of assembling protein clusters, underlines the importance of signaling complexes [Bray, 1998].
Although the method of *k-core communities* is concipated to find such complete subgraphs and its aggregations into *communities*, only signal transduction or transcription factor proteins are found to comprise such large and highly

connected regions which are embedded in a hierarchical core to periphery architecture.
Another interesting result is that proteins of the inner core are enriched in SH2 and SH3 domains, protein motifs that serve as docking sites for phosphorylated tyrosines and relay the signal onward [Alberts, 2002]. SH2 and SH3 domains have evolved to form multiprotein *scaffolds* and to regulate advanced biological functions by establishing switches, or interlink pathways to intricate pathway networks (*cross-talk*). Phospho-dependent domains provide an adaptable and dynamic mechanism for regulating cellular functions [Nash & Pawson,].
Proteins of the signaling core are involved in EGFR and mitosis signaling. If the signaling adapter is active in permanent association as one complex or as a scaffold is not examined yet, measuring complexes experimentally is still a difficult task. Similar to hubs, proteins having more than 30-50 binding partners (note that there is no definition for hubs), especially date-hubs [Han et al., 2004] being *nodes* with multiple impact, this relay station or integration point might provide the cell also with an extreme powerful property.
It is well known that protein domains are a highly variable concept in nature to effectively respond to cellular changes. Not any signaling (transcription) protein is associated to the hierarchical shells ($\sim$one fourth of all signal transduction proteins in the presented network), but the observed structure might represent a global *biological* backbone. Nevertheless, the *CPM* method might be too inflexible to cover more associated proteins by topology.
Two of the top-hubs, TP53 and SLC2A4, are not involved in the signaling or transcription shells, but *nodes* with *degree* higher than 100. The observed nested structures may serve as integration points on a higher level: hub proteins (esp. date-hubs; e.g. TP53) mediate many functions- the observed *hub-complexes* should have an incomparable larger potentiality to mediate and regulate biological functions.
Preliminary analysis whether a similar hierarchical structure is also present in PPI networks from other species (*Drosophila melanogaster* and *Sacharomyces cerevisiae*) show a reminiscent (but far less pronounced) organisation for fly, potentially due to multicellularity, but not for yeast. Multicellular organisms might need a more sophisticated architecture for communication. Very interestingly, the two proteins that take part in most *communities* in the *Drosophila* network are CG10079 (EGFR) and CG9375 (RAS85D), proteins important for mitosis (data not shown). Regarding the *Sacharomyces cerevisiae* network, cdc28 which is the catalytic subunit of the main cell cycle cyclin-dependent kinase CDK [Lew, 1997], is most central, followed by typical *communities* of the transcription machinery. This first result suggests that mitogenic signaling is a very complex and central function that is mirrored by an intricate topology.
The *CPM* method is very rigid as the algorithm allows only unions of full-connected subnetworks. To test the discovery of signaling and transcription hub integration points (cores and shells) with a less rigid method, a *k-core* decomposition is applied (different from *k-core communities*) and found as well a tightly interlinked hub-core.

5.4.2 The skeleton

The Minimal Spanning Tree based on highest *betweenness centrality*, called *skeleton* or communication kernel, is calculated here and analysed for a large human PPI network. This study shows that the *skeleton* of the human PPI network is also *scale-free*. Additionally, it is observed that proteins highly interwired on the original network reflect all signaling or transcription proteins. To get new insights into the organisation of biological communication, nodes having the highest number of *shortcuts* comparing the *skeleton* with the original network, are examined in detail to find proteins that are most important in communication processes.
Goh et al. stated that *shortcuts* are present mainly inside the *modules* and the interconnections between *modules* are accomplished largely through the *skeleton*. Referring to the human network, many shortcuts are present for signal transduction and transcription factors, which are two separated *modules* and compartments. As protein *modules* are densely clustered regions, other *modules* will presumably have also a lot of *shortcuts*. A first survey showed that the Histone Deacetylase complex, the Tata-binding protein and associated TAFs as well as integrin complexes are among the highest *shortcut* proteins (see chapter 4 for more details of dense complexes) supporting the hypothesis of Goh et al. [Goh et al., 2003]. The presented method for the calculation of *shortcuts* may function as a new method for depicting *modules* (further studies underway).
Kim et al. [Kim et al., 2004] suggested that the remaining shortcuts are responsible for the clustering property. However, the skeleton is based on the *betweenness centrality* which calculates the frequency of lying on shortest paths. The *BC* correlates highly with the degree, according to this, Newman (and references therein [Newman, 2005]) questions the use of the *BC*. However, there are a small number of *nodes* where *degree* and *betweenness* is very different so that the measurement identifies these *nodes* which might have a nontrivial impact on the network (see also chapter 2 interpreting *high betweenness- low connectivity nodes* as potential gatekeeper for intracellular trafficking aspects).
Which PPIs are alternatives or *shortcuts* is a difficult question from the biological perspective. The information has to be passed along a communication channel between or to *modules*, but this path is not necessarily the shortest, it will rather be the fastest or safest. Whether the MST *skeleton* is indeed the biological communication backbone is still in question. The observed shells are highly-redundant from a global perspective and the participating/ constituting proteins have a high number of *shortcuts*. From this perspective, the skeleton finds well clustered *regions* which is helpful in determining modules here. As a new attempt to overcome the limitations of the *shortest paths betweenness*, Newman defined a *random-walk betweenness centrality* where information spread is calculated along *all* paths instead of shortest paths. Although the *random-walk betweenness* tends to be higher for *nodes* with high betweenness centrality and hence high *degree*, a few *nodes* have very different values. As a further point for research, an MST based on *random walk betweenness* would be interesting, especially for the highly entangled cell communication processes.

The distribution of the *shortcut lengths* on the *skeleton* distinguishes *scale-free* networks into two types [Kim et al., 2004]. It is shown here that the human network (*skeleton* and original network) is from the longer- loop dominant structure. By reducing about three fourth of PPIs the average distance (and maximal distance) is only moderately increased by one third. This finding shows that the *skeleton* retains the shortest paths *edges* (by construction). Although helpful, biology might not act on this paths. A first interpretation does not support that the specific *edges* defining the *skeleton* are *the* protein interactions that are of most importance for exchanging information. In contrary, *HBLC nodes* (not *edges*; see chapter 3) might represent important connections.
The highest *relative* decrease of *shortcuts* (opposed to the *absolute* as examined here) affects mainly proteins with less than *degree* 60 and represent another interesting group of proteins for further studies. If the *community* structure is widely disrupted when deleting the MST in the giant component might give new insights for the role and feasibility of *edge betweenness* calculations.
The scale-free structure is maintained on the *skeleton*, because top-hubs do not change in *degree* relation. Interestingly, two groups of high *degree* proteins are observed here: the first group consists of proteins with a low number of *shortcuts* (e.g. SLC2A4, TP53, EWSR1), the second consists of proteins with many shortcuts (e.g. FOS, PTPN11, NCOR2 etc.) and is highly enriched in signaling proteins and transcription factors. Recently, hubs are classified into party-hubs that interact with their partners simultaneously and date-hubs which bind to their interactors at different time and subcellular locations [Han et al., 2004]. If the first group rather represents date-hubs bringing *modules* or *communities* together and the second forms party-hubs is an interesting idea for further analysis.
It is to mention that the presence of *rich-clubs* is a very new feature and the correct measurement is still in discussion [Colliza et al., 2006] [Luo et al., 2007]. Colizza et al. showed the *rich-club* structure to be absent in a yeast protein network compared to a scientific collaboration (social) network by applying an appropriate null-hypothesis. However, Zhou et al. [Luo et al., 2007] use a slightly other method stating that a *rich-club is* present in yeast. The human network presented here exhibits a *rich-club* especially for cell communication. Regarding the observed cell communication shells as *rich-clubs* combined to the low dissortative property, this topic is a highly interesting task for further analysis as one main result of this thesis.

5.4.3 Degree-correlation

Maslov and Sneppen showed that yeast networks feature a dissortative degree-degree correlation, denoting that highly connected proteins predominantly interact with less connected proteins[Maslov & Sneppen, 2002] and thus *cross-talk* as well as error propagation of mutations is minimized. However, it is found here that the dissortative degree correlation is very weakly pronounced in the human network. This is mostly due to the complex wiring of hub-

proteins into hub-networks for signal transduction and transcription factors. Both functions are known to need a high variety of *cross-talk*. Stelzl et al. [Stelzl et al., 2005] found as well a low correlation for a smaller human network and inferred that this feature might be less pronounced in multicellular organisms. Schueler et al. [Schüler, 2006] examined three eukaryotic networks (human (this network), fly, yeast) and compared the results to Maslov and Sneppen [Maslov & Sneppen, 2002] and Stelzl et al. All three representative networks exhibit low correlations, probably a sampling effect of the other published results by yeast two-hybrid screens. A high variation of the *degree* correlation is observed for proteins with a *degree* higher than 60. Interestingly, proteins of the inner core and the next layer have mostly degrees higher than 60. These proteins form a tightly interlinked subnetwork and thus the distribution may be more scattered in this *degree* segment.
In conclusion, many top-hub proteins constitute either a dense signaling or transcription integration point embedded in less but yet high degree hierarchical shells. The different shells are hierarchical, but not scale-free, over about three to four orders of magnitude. These two very central functions exchange information via topologically revealed shuttling proteins mediating between the cytoplasm and the nucleus.
The underlying theoretical communication backbone (*skeleton*) has the same scale-free property as the original network and proteins which *edges* that are highly excluded, representing redundant shortcuts, overlap largely with proteins from the inner cores.
The examined degree-degree correlation is only slightly dissortative, many of the top-hub proteins (degree higher than 60) interact with each other to enhance *cross-talk* for intracellular communication.

In summary, top-hub integration points are observed from three different levels and with three different methods: *regionally* by the *CPM* method, *globally* by the *skeleton* and locally by the local scattering of the *degree* correlation and the presence of a *rich-club*. The structure of the human PPI network is dominated by complex communication processes in the cytoplasm and the nucleus. Further studies may focus on the question whether other cellular networks (transcriptome, metabolome) or non-biological networks from society and technology feature a similar structure to understand and distinguish global design principles for complex networks. The presented topological features may be of high value for the understanding of cell communication and the EGFR-mediated pathways for cell proliferation and survival which are central biological processes that are disturbed in cancer.
Recently, a very new feature for complex networks was discovered, the property of self-similarity [Song et al., 2005], [Song et al., 2006], [Strogatz, 2005].
Further studies are underway to examine if the human network is also fractal and self-similar as shown for other biological networks. When fractality is accompanied with a high repulsion, how is the finding of *rich-clubs* in accordance with fractality and modularity? Does the potential feature of self-similarity yield a new concept to decipher functional *modules* by pure topological aspects?

Abstract

Abstract: Topological Aspects of the Human Protein Interaction Network

It is currently widely accepted that the understanding of complex cell functions depends on an integrated network theoretical approach and not on an isolated view of the different molecular agents. Aim of this thesis was the examination of topological properties that mirror known biological aspects by depicting the human protein network with methods from graph- and network theory.
The presented network is a partial human interactome of 9222 proteins and 36324 interactions, consisting of single interactions reliably extracted from peer-reviewed scientific publications. In general, one can focus on intra- or intermodular characteristics, where a functional module is defined as "a discrete entity whose function is separable from those of other modules".
It is found that the presented human network is also scale-free and hierarchically organised, as shown for yeast networks before. The interactome also exhibits proteins with high betweenness and low connectivity which are biologically analysed and interpreted here as shuttling proteins between organelles (e.g. ER to Golgi, internal ER protein translocation, peroxisomal import, nuclear pores import/export) for the first time. As an optimisation for finding proteins that connect modules, a new method is developed here based on proteins located between highly clustered regions, rather than regarding highly connected regions. As a proof of principle, the "Mediator complex" is found in first place, the prime example for a connector complex. Focusing on intramodular aspects, the measurement of k-clique communities discriminates overlapping modules very well. Twenty of the largest identified modules are analysed in detail and annotated to known biological structures (e.g. proteasome, the NFκB-, TGF-β complex). Additionally, two large and highly interconnected modules for signal transducer and transcription factor proteins are revealed, separated by known shuttling proteins. These proteins yield also the highest number of redundant shortcuts (by calculating the skeleton), exhibit the highest numbers of interactions and might constitute highly interconnected but spatially separated rich-clubs either for signal transduction or for transcription factors. This design principle allows manifold regulatory events for signal transduction and enables a high diversity of transcription events in the nucleus by a limited set of proteins.
Altogether, biological aspects are mirrored by pure topological features, leading to a new view and to new methods that assist the annotation of proteins to biological functions, structures and subcellular localisations. As the human protein network is one of the most complex networks at all, these results will be fruitful for other fields of network theory and will help understanding complex network functions in general.

Zusammenfassung: Topologische Aspekte des humanen Proteininteraktionsnetzwerkes

Es ist mittlerweile anerkannt, dass das Verständnis komplexer Zellfunktionen einen integrierten netzwerktheoretischen Ansatz und nicht die isolierte Sichtweise der Funktion einzelner Faktoren bedingt. Ziel dieser Arbeit war es, topologische Methoden zu überprüfen und zu entwickeln, die bekannte biologische Aspekte widerspiegeln, indem das Netzwerk mit graphen- und netzwerktheoretischen Methoden dargestellt wird.
Das hier vorgestellte Netzwerk ist ein partielles menschliches Interaktom mit 9222 Proteinen und 36324 Interaktionen, bestehend aus einzelnen Interaktionen, welche aus begutachteten wissenschaftlichen Veröffentlichungen extrahiert wurden. Generell kann man auf intra- oder intermodulare Charakteristiken fokussieren, wobei ein funktionelles Modul als "eine diskrete Einheit dessen Funktion von der anderer Module verschieden ist" definiert wird.
Es wurde gefunden, dass das untersuchte menschliche Netzwerk skaleninvariant und hierarchisch organisiert ist, wie zuvor für Hefenetzwerke gezeigt wurde. Das Interaktom weist Proteine mit einer hohen "betweenness" und einer niedrigen Konnektionierung auf, welche hier zum ersten Mal biologisch als "Shuttlingproteine" zwischen Organellen analysiert und interpretiert wurden (z.B. ER zu Golgi, interne ER-Proteintranslokation, peroxisomaler Transport, Import/Export durch nukleäre Poren). Als Optimierung für das Auffinden von Proteinen, die Module konnektionieren, wurde eine neue Methode entwickelt, die auf Proteine basiert, die zwischen hochvernetzten und nicht hochkonnektionierten Regionen lokalisiert sind. Als Nachweis der Methode, wurde der "Mediatorkomplex" an erster Stelle gefunden, das beste Beispiel für einen Konnektorkomplex. Auf intramodulare Aspekte bezogen, unterscheidet die Methode der "k-clique communities" überlappende Module. Zwanzig der größten identifizierten Module sind detailliert analysiert und bekannten biologischen Strukturen zugeordnet worden (z.B. Proteasom, NFκB-, TGF-β Komplex). Zusätzlich wurden zwei große und hochvernetzte Module für Signaltransduktions- und Transkriptionsproteine gefunden, die durch bekannte "Shuttlingproteine" getrennt sind. Diese Proteine weisen auch die höchste Anzahl an redundanten "shortcuts" auf (durch Berechnung des "skeletons"), haben die höchste Anzahl an Interaktionen und könnten hochvernetzte aber räumlich getrennte "richclubs", entweder für Signaltransduktions- oder Transkriptionsfaktoren darstellen. Dieses Designprinzip ermöglicht vielseitige regulatorische Möglichkeiten, um Signale zu translozieren und eine hohe Diversität an Transkriptionsmöglichkeiten im Nukleus mit einer geringen Anzahl von Proteinen umzusetzen.
Zusammenfassend lässt sich sagen, dass biologische Aspekte über bloße topologische Eigenschaften widergespiegelt werden, dies führt zu einer neuen Sichtweise und zu neuen Methoden, die bei der Zuordnung von Proteinen zu biologischen Funktionen, Strukturen und subzellulärer Lokalisation assistieren. Da das menschliche Proteinnetzwerk eines der komplexesten Netwerke überhaupt ist, werden diese Ergebnisse gewinnbringend für andere Bereiche der Netzwerktheorie sein und dabei helfen komplexe Netzwerke generell zu verstehen.

Bibliography

[Adamcsek et al., 2006] Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., & Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8), 1021–1023.

[Alberts, 2002] Alberts (2002). *Molecular Biology of the Cell.* Garland.

[Alberts, 1998] Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3), 291–294.

[Alfarano et al., 2005] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P., Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F. F., & Hogue, C. W. V. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*, 33(Database issue), D418–D424.

[Andreev et al., 2001] Andreev, J., Galisteo, M. L., Kranenburg, O., Logan, S. K., Chiu, E. S., Okigaki, M., Cary, L. A., Moolenaar, W. H., & Schlessinger, J. (2001). Src and Pyk2 mediate G-protein-coupled receptor activation of epidermal growth factor receptor (EGFR) but are not required for coupling to the mitogen-activated protein (MAP) kinase signaling cascade. *J Biol Chem*, 276(23), 20130–20135.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S.,

Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1), 25–29.

[Barabasi, 2005] Barabasi, A.-L. (2005). Taming Complexity. *Nature Physics*, 1, 68–70.

[Barabási & Oltvai, 2004] Barabási, A.-L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2), 101–113.

[Batagelj & Mrvar, 1998] Batagelj, V. & Mrvar, A. (1998). Pajek – program for large network analysis.

[Beggs, 2005] Beggs, J. D. (2005). Lsm proteins and RNA processing. *Biochem Soc Trans*, 33(Pt 3), 433–438.

[Belakavadi & Fondell, 2006] Belakavadi, M. & Fondell, J. D. (2006). Role of the mediator complex in nuclear hormone receptor signaling. *Rev Physiol Biochem Pharmacol*, 156, 23–43.

[Blinov et al., 2006] Blinov, M. L., Faeder, J. R., Goldstein, B., & Hlavacek, W. S. (2006). A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *Biosystems*, 83(2-3), 136–151.

[Bray, 1998] Bray, D. (1998). Signaling complexes: biophysical constraints on intracellular communication. *Annu Rev Biophys Biomol Struct*, 27, 59–75.

[Büttner et al., 2006] Büttner, K., Wenig, K., & Hopfner, K.-P. (2006). The exosome: a macromolecular cage for controlled RNA degradation. *Mol Microbiol*, 61(6), 1372–1379.

[Burley & Roeder, 1998] Burley, S. K. & Roeder, R. G. (1998). TATA box mimicry by TFIID: autoinhibition of pol II transcription. *Cell*, 94(5), 551–553.

[Chen & Yuan, 2006] Chen, J. & Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 22(18), 2283–2290.

[Colliza et al., 2006] Colliza, Flammini, Serrano, & Vespignani (2006). Detecting rich-club ordering in complex networks. *Nature Physics*, 2, 110–115.

[Derényi et al., 2005] Derényi, I., Palla, G., & Vicsek, T. (2005). Clique percolation in random networks. *Phys Rev Lett*, 94(16), 160202.

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

[Eckert et al., 2004] Eckert, R. L., Broome, A.-M., Ruse, M., Robinson, N., Ryan, D., & Lee, K. (2004). S100 proteins in the epidermis. *J Invest Dermatol*, 123(1), 23–33.

[Feng et al., 2001] Feng, Y., Longo, D. L., & Ferris, D. K. (2001). Polo-like kinase interacts with proteasomes and regulates their activity. *Cell Growth Differ*, 12(1), 29–37.

[Fernandez et al., 2004] Fernandez, C. F., Pannone, B. K., Chen, X., Fuchs, G., & Wolin, S. L. (2004). An Lsm2-Lsm7 complex in Saccharomyces cerevisiae associates with the small nucleolar RNA snR5. *Mol Biol Cell*, 15(6), 2842–2852.

[Futschik et al., 2007] Futschik, M. E., Chaurasia, G., Tschaut, A., Russ, J., Babu, M. M., & Herzel, H. (2007). Functional and transcriptional coherency of modules in the human protein interaction network. *Journal of Integrative Bioinformatics.*, 4(3), 76.

[Gallinari et al., 2007] Gallinari, P., Marco, S. D., Jones, P., Pallaoro, M., & Steinkühler, C. (2007). HDACs, histone deacetylation and gene transcription: from molecular biology to cancer therapeutics. *Cell Res*, 17(3), 195–211.

[Gavin et al., 2002] Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868), 141–147.

[Giot et al., 2003] Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., & Rothberg, J. M. (2003). A protein interaction map of Drosophila melanogaster. *Science*, 302(5651), 1727–1736.

[Goh et al., 2003] Goh, K.-I., Oh, E., Kahng, B., & Kim, D. (2003). Betweenness centrality correlation in social networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(1 Pt 2), 017101.

[Hamsen, 2005] Hamsen, G. (2005). Development of disease-specific data environments using literature-based information an protein interaction data. Master's thesis, University of Applied Sciences of Gelsenkirchen.

[Han et al., 2004] Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P., & Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995), 88–93.

[Hartwell et al., 1999] Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl), C47–C52.

[Hermjakob et al., 2004] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., & Apweiler, R. (2004). The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2), 177–183.

[Hoffjan & Stemmler, 2007] Hoffjan, S. & Stemmler, S. (2007). On the role of the epidermal differentiation complex in ichthyosis vulgaris, atopic dermatitis and psoriasis. *Br J Dermatol*, 157(3), 441–449.

[Houseley et al., 2006] Houseley, J., LaCava, J., & Tollervey, D. (2006). RNA-quality control by the exosome. *Nat Rev Mol Cell Biol*, 7(7), 529–539.

[Hwang et al., 2006] Hwang, Cho, Zhang, & Ramanathan (2006). Bridging Centrality: Identifying Bridging Nodes In Scale-free Networks. *KDD'06 August 2023,*, Philadelphia, PA, USA.

[Ito et al., 2001] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8), 4569–4574.

[Jorissen et al., 2003] Jorissen, R. N., Walker, F., Pouliot, N., Garrett, T. P. J., Ward, C. W., & Burgess, A. W. (2003). Epidermal growth factor receptor: mechanisms of activation and signalling. *Exp Cell Res*, 284(1), 31–53.

[Joy et al., 2005] Joy, M. P., Brock, A., Ingber, D. E., & Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *J Biomed Biotechnol*, 2005(2), 96–103.

[Kanehisa et al., 2006] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue), D354–D357.

[Kim et al., 2004] Kim, D.-H., Noh, J. D., & Jeong, H. (2004). Scale-free trees: the skeletons of complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(4 Pt 2), 046126.

[Krauss, 2003] Krauss (2003). *Biochemistry of Signal Transduction and Regulation.* Wiley-VCH, Weinheim.

[Lew, 1997] Lew (1997). *The Molecular and Cellular Biology of the Yeast Saccharomyces: Cell Cycle and Cell Biology.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

[Loh & Hong, 2004] Loh, E. & Hong, W. (2004). The binary interacting network of the conserved oligomeric Golgi tethering complex. *J Biol Chem*, 279(23), 24640–24648.

[Luo et al., 2007] Luo, F., Yang, Y., Chen, C.-F., Chang, R., Zhou, J., & Scheuermann, R. H. (2007). Modular organization of protein interaction networks. *Bioinformatics*, 23(2), 207–214.

[Machida & Mayer, 2005] Machida, K. & Mayer, B. J. (2005). The SH2 domain: versatile signaling module and pharmaceutical target. *Biochim Biophys Acta*, 1747(1), 1–25.

[Maslov & Sneppen, 2002] Maslov, S. & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, 296(5569), 910–913.

[Minucci & Pelicci, 2006] Minucci, S. & Pelicci, P. G. (2006). Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. *Nat Rev Cancer*, 6(1), 38–51.

[Nakao, 1990] Nakao, K. (1990). Distribution of measures of centrality: Enumerated distributions of Freeman's graph centrality measures. *Connections*, 13(3), 10–22.

[Nash & Pawson,] Nash, P. & Pawson, T. T Phospho-dependent Protein Interaction Domains in Signal Transduction. http://www.cellsignal.com/reference/domain/index.jsp.

[NetPro, 2005] NetPro (2005). http://www.molecularconnections.com/products.html.

[Newman, 2002] Newman (2002). *Random graphs as models of networks. In Handbook of Graphs and Network.* Wiley-VHC, Berlin.

[Newman, 2005] Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 1, 39–54.

[Nilsen, 2003] Nilsen, T. W. (2003). The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, 25(12), 1147–1149.

[Oda et al., 2005] Oda, K., Matsuoka, Y., Funahashi, A., & Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol*, 1, 2005.0010.

[Palla et al., 2005] Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043), 814–818.

[Park & Choi, 2006] Park, C.-J. & Choi, B.-S. (2006). The protein shuffle. Sequential interactions among components of the human nucleotide excision repair pathway. *FEBS J*, 273(8), 1600–1608.

[Peri et al., 2003] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G. N., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., & Pandey, A. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10), 2363–2371.

[Przulj, 2005] Przulj (2005). *Graph Theory-Analysis of Protein-Protein Interactions in Knowledge Discovery in Proteomics*. CRC Press Inc.

[Ravasz & Barabási, 2003] Ravasz, E. & Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(2 Pt 2), 026112.

[Reich & Liu, 2006] Reich, N. C. & Liu, L. (2006). Tracking STAT nuclear traffic. *Nat Rev Immunol*, 6(8), 602–612.

[Richardson & Zundel, 2005] Richardson, K. S. & Zundel, W. (2005). The emerging role of the COP9 signalosome in cancer. *Mol Cancer Res*, 3(12), 645–653.

[Sala-Valdés et al., 2006] Sala-Valdés, M., Ursa, A., Charrin, S., Rubinstein, E., Hemler, M. E., Sánchez-Madrid, F., & Yáñez-Mó, M. (2006). EWI-2 and EWI-F link the tetraspanin web to the actin cytoskeleton through their direct association with ezrin-radixin-moesin proteins. *J Biol Chem*, 281(28), 19665–19675.

[Schaner & Gumucio, 2005] Schaner, P. E. & Gumucio, D. L. (2005). Familial Mediterranean fever in the post-genomic era: how an ancient disease is providing new insights into inflammatory pathways. *Curr Drug Targets Inflamm Allergy*, 4(1), 67–76.

[Schüler, 2006] Schüler, A. (2006). Development of a software for the analysis of cellular networks and computational analysis of the human interactome. Bachelor Thesis.

[Schüler et al., 2005a] Schüler, A., Brinck, H., Lutter, P., Schmitt, E., Jonuleit, H., & Wiebringhaus, T. (2005a). Proteins of the epidermal differentiation complex (EDC) are differentially expressed in CD4(+)CD25(+) regulatory T-cells. In *German Conference on Bioinformatics 2005*.

[Schüler et al., 2005b] Schüler, A., Sonneborn, B., Perrey, S., Brinck, H., & Wiebringhaus, T. (2005b). Modeling human signaling pathways for complex diseases. In *German Conference on Bioinformatics 2005*.

[Schwaebe, 2005] Schwaebe, A. (2005). Konzeption eines Assistenzsystems zur Extraktion von Informationen über Protein-Protein Wechselwirkungen aus Fachliteratur. Diploma Thesis.

[Scott et al., 2006] Scott, J., Ideker, T., Karp, R. M., & Sharan, R. (2006). Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol*, 13(2), 133–144.

[Sedgewick & Wyk, 2002] Sedgewick, R. & Wyk, C. V. (2002). *Algorithms in C++, Part 5 Graph Algorithms.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

[Sekiguchi et al., 2004] Sekiguchi, T., Todaka, Y., Wang, Y., Hirose, E., Nakashima, N., & Nishimoto, T. (2004). A novel human nucleolar protein, Nop132, binds to the G proteins, RRAG A/C/D. *J Biol Chem*, 279(9), 8343–8350.

[Song et al., 2006] Song, C., Havlin, S., & Makse, H. (2006). Origins of fractality in the growth of complex networks. *Nature Physics*, 2, 275–81.

[Song et al., 2005] Song, C., Havlin, S., & Makse, H. A. (2005). Self-similarity of complex networks. *Nature*, 433(7024), 392–395.

[Sonneborn, 2005] Sonneborn, B. (2005). Modellierung von Pfaden im humanen Protein-Protein-Interaktionsnetzwerk. Diploma Thesis.

[Sonneborn et al., 2005] Sonneborn, B., Schüler, A., Perrey, S., Brinck, H., & Wiebringhaus, T. (2005). Modeling human signaling pathways for complex diseases. In *Therapeutic Applications of Computational Biology (EBI Hinxton, 2005) Conference Programme p.21.*

[Stelzl et al., 2005] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., & Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6), 957–968.

[Stewart, 2007] Stewart, M. (2007). Molecular mechanism of the nuclear protein import cycle. *Nat Rev Mol Cell Biol*, 8(3), 195–208.

[Strogatz, 2005] Strogatz, S. H. (2005). Complex systems: Romanesque networks. *Nature*, 433(7024), 365–366.

[Uetz et al., 2000] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., & Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770), 623–627.

[Ungar et al., 2005] Ungar, D., Oka, T., Vasile, E., Krieger, M., & Hughson, F. M. (2005). Subunit architecture of the conserved oligomeric Golgi complex. *J Biol Chem*, 280(38), 32729–32735.

[Urbanczyk, 2006] Urbanczyk, C. (2006). Analyse von Gen- und Proteinsynonymen. Bachelor Thesis.

[Watts & Strogatz, 1998] Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442.

[Welch, 1999] Welch, M. D. (1999). The world according to Arp: regulation of actin nucleation by the Arp2/3 complex. *Trends Cell Biol*, 9(11), 423–427.

[Wiebringhaus et al., 2005a] Wiebringhaus, T., Brinck, H., Lutter, P., Schmitt, E., & Jonuleit, H. (2005a). Proteins of the epidermal differentiation complex (EDC) are differentially expressed in CD4(+)CD25(+) regulatory T-cells. In *HUPO 4th Annual World Congress, Molecular & Cellular Proteomics Vol.4 no.8, p. 58.*

[Wiebringhaus et al., 2005b] Wiebringhaus, T., Hamsen, G., Hambsch, B., Brinck, H., Schulenborg, T., May, C., Schmidt, O., Meyer, H., & Marcus, K. (2005b). Differential Proteome Analyses and Protein Interaction Networks of disease- related mouse models for Alzheimer's and Parkinson's Disease. In *NGFN2 Meeting 2005 and HUPO Human Brain Proteome Meeting.*

[Wiebringhaus et al., 2004a] Wiebringhaus, T., Perrey, S., & Brinck, H. (2004a). Comparative Analysis and Text Mining of a functional genomic cluster in a susceptibility region for Psoriasis and Atopic Dermatitis. In *German Conference on Bioinformatics 2004, Proceedings p. 54-55.*

[Wiebringhaus et al., 2004b] Wiebringhaus, T., Schwaebe, A., Centler, F., & Brinck, H. (2004b). Human curated and assisted extraction of protein interactions for discovering disease-relevant pathways. In *German Conference on Bioinformatics 2004, Proceedings p. 56-57.*

[Wiiger & Prydz, 2004] Wiiger, M. T. & Prydz, H. (2004). The epidermal growth factor receptor (EGFR) and proline rich tyrosine kinase 2 (PYK2) are involved in tissue factor dependent factor VIIa signalling in HaCaT cells. *Thromb Haemost*, 92(1), 13–22.

[Wolf et al., 2003] Wolf, D. A., Zhou, C., & Wee, S. (2003). The COP9 signalosome: an assembly and maintenance platform for cullin ubiquitin ligases? *Nat Cell Biol*, 5(12), 1029–1033.

[Wu et al., 2001] Wu, C. J., O'Rourke, D. M., Feng, G. S., Johnson, G. R., Wang, Q., & Greene, M. I. (2001). The tyrosine phosphatase SHP-2 is required for mediating phosphatidylinositol 3-kinase/Akt activation by growth factors. *Oncogene*, 20(42), 6018–6025.

[Xu, 2006] Xu, L. (2006). Regulation of Smad activities. *Biochim Biophys Acta*, 1759(11-12), 503–513.

[Yang et al., 2006] Yang, X. H., Kovalenko, O. V., Kolesnikova, T. V., Andzelm, M. M., Rubinstein, E., Strominger, J. L., & Hemler, M. E. (2006). Contrasting effects of EWI proteins, integrins, and protein palmitoylation on cell surface CD9 organization. *J Biol Chem*, 281(18), 12976–12985.

[Yu et al., 2006] Yu, H., Paccanaro, A., Trifonov, V., & Gerstein, M. (2006). Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7), 823–829.

[Zhang et al., 2006] Zhang, S., Ning, X., & Zhang, X.-S. (2006). Identification of functional modules in a PPI network by clique percolation clustering. *Comput Biol Chem*, 30(6), 445–451.

Supplementary

Biological Annotation and Local Structure of Communities

This part shows citations from the NCBI websites, which are solely used here to demonstrate and show the annotation of protein complexes to biological processes or structures.
For readability reasons, some sections are therefore shortened or extended with own words (without quoting) and therefore vary compared to the original sources. Please refer especially to the NCBI Gene websites when not declared otherwise.

The following sections show manually annotated *communities*, where k reflects the *k-core*, V the number of proteins, E the number of *edges, c* the *community* number and CC_{n} the density of the *community.*

Tata-binding proteins (TBPs) and associated factors (TAFs)

More than 70 polypeptides are required for transcription initiation by RNA polymerase II. An important regulator for positioning the polymerase II is transcription factor IID (TFIID). TFIID binds to the core promoter and acts as a scaffold for the remaining proteins, it is composed of the TATA-binding protein (TBP) and evolutionary conserved TBP-associated factors (TAFs). TAFs have a variety of functions, including basal transcription, coactivation, promoter recognition or transcription initiation (NCBI Gene, see also [Burley & Roeder, 1998]).

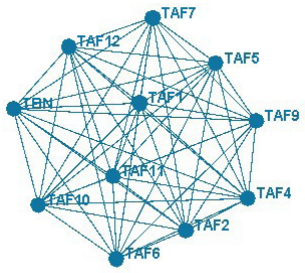


Figure 15: Eleven TAF proteins as a full-connected subgraph (clique) comprising 55 *edges*, constitutes the largest *k-core community*.

Naturally the complex occurs also at lower k (fig. 16). At k6, functional relevant proteins are included into the *community* showing the feasibility of the method for biological applications. TAF15 and the TATA-binding protein TBP are included at k6 (fig. 16) and the Cleavage and Polyadenylation specific factor 11 CPSF1 is included at k5 (fig. 17). The well-known TFIID complex is thus reconstituted by pure topological aspects (proof-of-principle).

Table 12: Tata-binding protein (TBP) and associated factors (TAFs) as a large full-connected clique with 55 interactions.

k	c	V	E	CC_{n}	Biological Activity	Reference
11	0	11	55	1	Tata-binding protein (TBP) and associated factors (TAFs)	[Burley & Roeder, 1998]
10	1	11	55	1		
9	1	11	55	1		
8	2	11	55	1		
7	5	11	55	1		
6	13	13	65	0,8		
5	41	14	69	0,72		

Regarding this complex, except TAF3, any TAF1-11 is interconnected (TAF8 is synonymous to TBN).

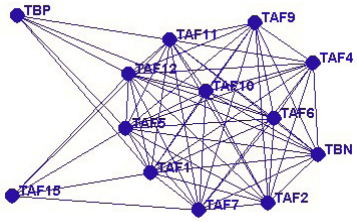


Figure 16: Thirteen TAF proteins at k=6. The method includes TBP and TAF15 which are reported as functionally relevant for the TAF complex.

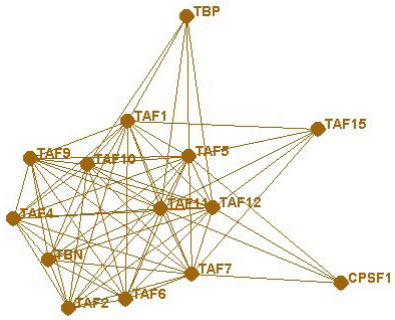


Figure 17: Fourteen TAF proteins at k=5 including cleavage and polyadenylation specific factor 11 (CPSF1).

Histone Deacetylase Complex (HDAC) (Sin3A)

Histone deacetylases (HDACs) enzymatic activity controls the acetylation state of the core histones. Acetylation of histones affects gene expression through its influence on chromatin conformation. HDACs intervene in a multitude of biological processes and are part of a multiprotein family in which each member has its specialized functions. Control of cell cycle progression, cell survival and differentiation are among the most important roles of HDACs. Since these processes are affected by malignant transformation, HDAC inhibitors were developed as antineoplastic drugs and are showing efficiency in cancer patients (taken and adapted from [Gallinari et al., 2007])(see fig. 18 and table 13)

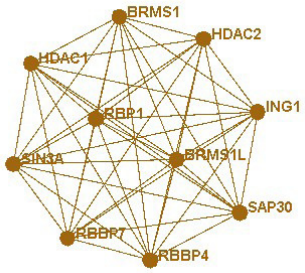


Figure 18: Histone Deacetylase Complex HDAC also known as Sin3A responsible for histone acetylation and control of gene expression for essential functions like the cell cycle, survival and progression.

Table 13: Histone Deacetylase Complex HDAC or Sin3A. The complex is a full-connected clique with 10 proteins and 45 interactions.

k	c	V	E	CC_{n}	Biological Activity	Reference
10	0	10	45	1	Histone Deacetylase Complex (HDAC), Sin3A	[Minucci & Pelicci, 2006]
9	0	10	45	1		
8	1	10	45	1		
7	2	10	45	1		

The Exosome Complex

The exosome, a large multisubunit complex with exoribonucleic activity, emerges as the central 3' RNA degradation and processing factor in eukaryotes and archaea. Recent functional and structural progress shows that the exosome is a macromolecular cage, where the nuclease active sites are situated in a central processing chamber. The emerging mechanism of exosome function suggests a strikingly parallel architectural concept to protein degradation by proteasomes (taken from [Büttner et al., 2006])(see fig. 19 and table 14)

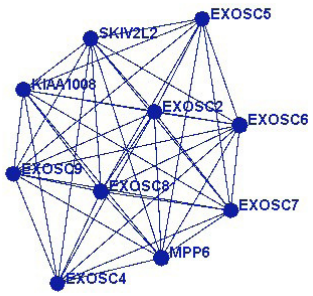


Figure 19: Ten proteins of the Exosome Complex important for RNA degradation.

Table 14: Exosome Complex important for RNA degradation.

k	c	V	E	CC_{n}	Biological Activity	Reference
9	2	10	44	0,97	Exosome complex	[Houseley et al., 2006]
8	3	10	44	0,97		
7	6	11	51	0,91		

COP9 Signalosome

The COP9 signalosome (CSN) is a highly conserved protein complex implicated in diverse biological functions that involve ubiquitin-mediated proteolysis. In the last several years, multiple lines of evidence have suggested that the CSN plays a significant role in the regulation of multiple cancers and could be an attractive target for therapeutic intervention. Deregulation of CSN subunit function can have a dramatic effect on diverse cellular functions, including the maintenance of DNA fidelity, cell cycle control, DNA repair, angiogenesis, and microenvironmental homeostasis that are critical for tumor development(taken and adapted from [Richardson & Zundel, 2005], [Wolf et al., 2003])(see fig. 20 and table 15).

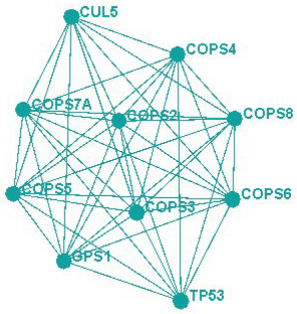


Figure 20: 10 proteins as part of the COP9 Signalosome important for ubiquitin-mediated proteolysis.

Table 15: The COP9 Signalosome.

k	c	V	E	CC_{n}	Biological Activity	Reference
9	3	10	44	0,97	COP9 Signalosome	[Wolf et al., 2003]
8	5	10	44	0,97		
7	8	10	44	0,97		

Actin-assembly

The coordination of cell shape change and locomotion requires that actin polymerization at the cell cortex is tightly controlled in response to both intracellular and extracellular signals. The Arp2/3 complex, an actin filament nucleating and organizing factor, appears to be a central player in the cellular control of actin assembly. This complex is located at the cell surface and is essential to cell shape and motility through actin assembly (taken and adapted from [Welch, 1999])(see fig. 21 and table 16).

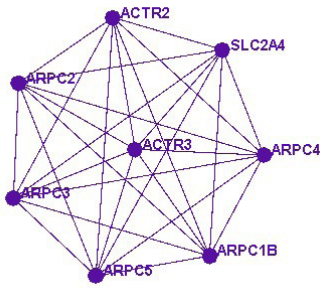


Figure 21: 7 proteins (except SLC2A4) of the ARP2/3 complex important for actin assembly.

Table 16: The ARP2/3 complex.

k	c	V	E	CC_{n}	Biological Activity	Reference
8	0	8	28	1	Actin-related protein complex ARP2/3	[Welch, 1999]
7	0	8	28	1		

Sm-like proteins (LSM)

Sm and LSM proteins are ubiquitous in eukaryotes and form complexes that interact with RNAs involved in almost every cellular process. Beggs et al. [Beggs, 2005] studied the Lsm proteins in the yeast *Saccharomyces cerevisiae*, identifying in the nucleus and cytoplasm distinct complexes that affect pre-mRNA splicing and degradation, small nucleolar RNA, tRNA processing, rRNA processing and mRNA degradation. These activities suggest RNA chaperone-like roles for Lsm proteins, affecting RNA-RNA and/ or RNA-protein interactions (taken and adapted from [Beggs, 2005]).
Dunn et al.[Dunn et al., 2005] found this complex with 8 proteins (LSM 1-7 and SMN1) as well by applying the *edge-betweenness centrality*, the method of *communities* examined here embeds more proteins into the biological LSM-context (LSM 8 in fig. 23) at lower k, demonstrating the advantages of this method. Note that the shown complete LSM-complex with 8 proteins has 94 further interaction partners with 259 protein interactions. Only one protein (LSM8) is included into the LSM-complex, using the *CPM* method. Interestingly, three complexes LSM1-7, LSM2-8 and LSM2-7 with differing functions are reported for yeast [Fernandez et al., 2004].

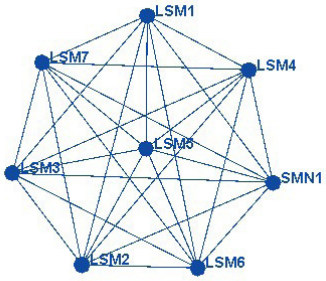


Figure 22: LSM1-7 complex as complete subgraph with 8 proteins (k8 c4) important for splicing and degradation of mRNAs.

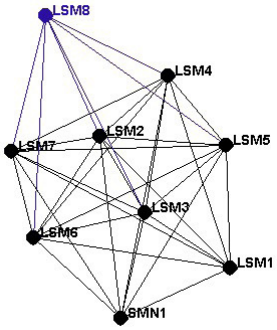


Figure 23: LSM8 is included into the subnetwork at lower k (k6 c15).

LSM8 binds to 31 further proteins in the presented network denoting different functions than complex formation alone. In yeast, the LSM2-LSM8 complex binds and stabilizes the spliceosomal U6 snRNA, whereas the LSM1-LSM7 complex functions in mRNA decay [Fernandez et al., 2004]. This result supposes a central role for LSM1 and LSM8, both do not interact with each other(not reported so far). Using an algorithm for enhancing the density of networks by adding only proteins that have at least 2 interaction partners in the original subgraph (method not shown), LSM8 links to the Spliceosome SNRPD-complex via SNRPD1 and to the RNA degradation Exosome complex via EXOSC6 and 10 while LSM1 does also link to the Exosome but *not* to the Spliceosome SNRPD-complex (data not shown).
This conjunction is not found by the method of *communities* alone, but by combining the two strategies. However, more interpretational effort is necessary to test these assumption.

Table 17: Sm-like proteins.

k	c	V	E	CC_{n}	Biological Activity	Reference
8	4	8	28	1	Sm-like proteins (LSM)	[Fernandez et al., 2004]
7	7	9	34	0,93		
6	15	9	34	0.93		

The Mediator complex

Mediator is an evolutionarily conserved multisubunit protein complex that plays a key role as a coactivator required for activation of RNA polymerase II transcription (see fig. 24) by DNA bound transcription factors. The complex functions by serving as a molecular bridge between DNA-bound transcriptional activators and the basal transcription apparatus (taken and adapted from [Belakavadi & Fondell, 2006])

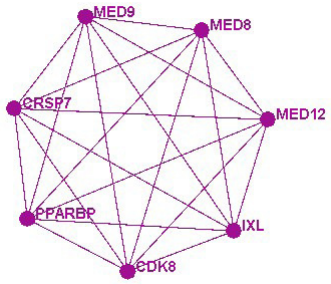


Figure 24: The mediator core complex.

Fig. 25 shows a dense region representing the RNA polymerase II which is responsible for synthesizing messenger RNA in eukaryotes.

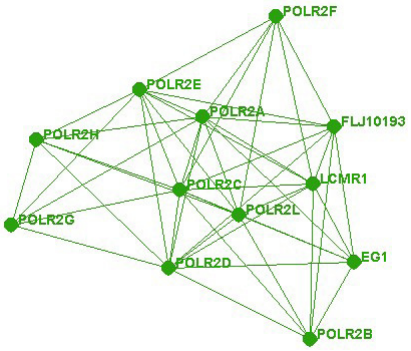


Figure 25: The RNA polymerase II complex at k6 c9; LCMR1(MED19).

The activation of gene transcription is a multistep process that is triggered by factors that recognize transcriptional enhancer sites in DNA. These factors work with co-activators to direct transcriptional initiation by the RNA polymerase II apparatus. The CRSP proteins shown here are subunits of the CRSP (cofactor required for SP1 activation) complex which is required for efficient activation by SP1. The mediator complex (LCMR1 (MED19) is central) brings together the CRSP proteins as coactivators with RNA polymerase II (POLR). (See fig. 26)

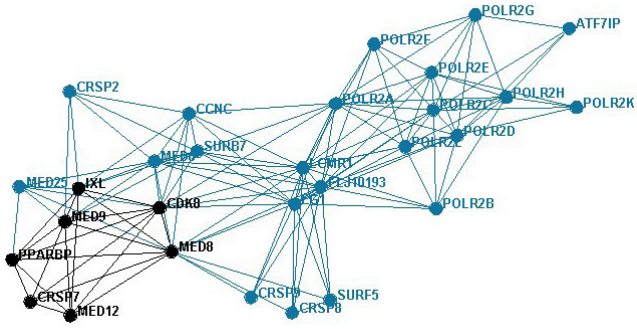


Figure 26: The mediator complex assembles the RNA polymerase II with coactivators (k=5, c25). Note that the CC_{n} is only 28%.

Table 18: The Mediator complex.

k	c	V	E	CC_{n}	Biological Activity	Reference
7	1	7	21	1	Mediator Complex	[Belakavadi & Fondell, 2006];
5	25	29	133	0,28		NCBI Gene

The thyroid hormone receptor associated proteins (THRAPs) are shown in conjunction with the mediator complex in figure 27. Note, that the TBP-associated factor TAF15 links the complexes to the Tata-binding protein (TBP) and thus to the TFIID-complex. Figure 28 shows the first step in transcription initiation.

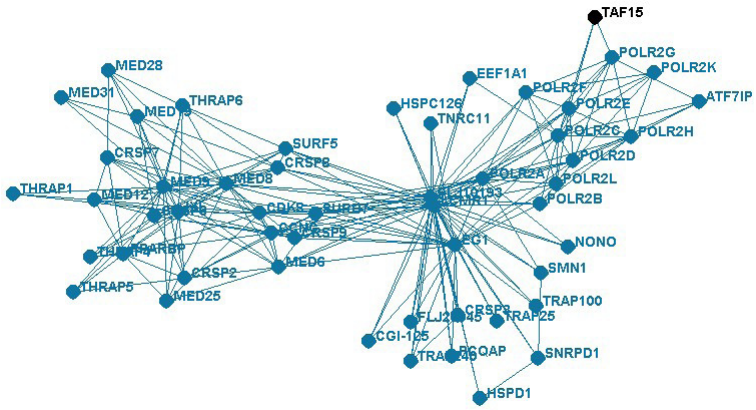


Figure 27: The essential steps in transcription initiation are reconstituted here by pure topological aspects(k4 c9).

Table 19: Transcription Initiation.

k	c	V	E	CC_{n}	Biological Activity	Reference
6	9	12	48	0,03	Polymerase	NCBI Gene
5	25	29	133	0,27	Transcription Initiation	NCBI Gene
4	9	52	226	0,14		

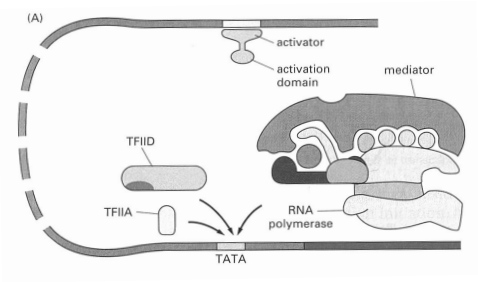


Figure 28: Alberts et al [Alberts, 2002].

The mediator complex is also recovered as a highly interconnecting complex using other methods in chapter 3.

Nucleotide Excision Repair (NER)

In mammalian cells, NER is a dynamic process in which a variety of proteins interact with one another, to carry out their functions. Xeroderma pigmentosum proteins are key players in several steps of the NER process, including DNA strand discrimination, repair complex formation, repair factor recruitment and other complex molecular interactions among NER factors in the context of DNA repair. Through these interactions, various types of bulky DNA strands can be recognized and repaired. (taken and adapted from [Park & Choi, 2006])

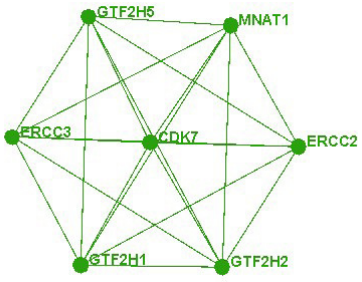


Figure 29: The Nucleotide Excision Repair complex (k7 c3).

Table 20: The Nucleotide Excision Repair complex.

k	c	V	E	CC_{n}	Biological Activity	Reference
7	3	7	21	1	DNA repair	[Park & Choi, 2006]

The Spliceosome and small nuclear Ribonucleoproteins (snRNPs)

Small nuclear ribonucleoprotein core proteins are required for pre-mRNA splicing and small nuclear ribonucleoprotein biogenesis. The process of splicing takes place in a massive ribonucleoprotein complex known as the spliceosome. Extensive studies have shown that splicing requires snRNPs, RNAs and many non-snRNP protein factors. It is revealed that the spliceosome is composed of as many as 300 distinct proteins and five RNAs, making it among the most complex macromolecular machines known (taken and adapted from [Nilsen, 2003]). Ten central proteins for the splicing process are found here. Further interpretational effort is needed to incorporate more proteins to the spliceosome. (See fig. 30 and table 21)

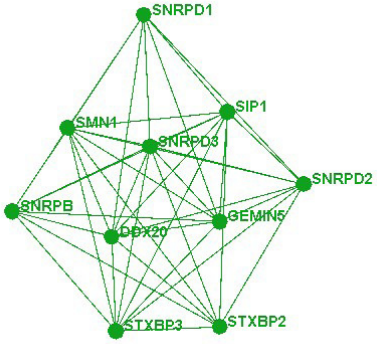


Figure 30: Proteins of the Spliceosome Complex (k7 c4).

Table 21: Proteins of the Spliceosome Complex.

k	c	V	E	CC_{n}	Biological Activity	Reference
7	4	10	41	0,89	Spliceosome	[Nilsen, 2003]

The Proteasome (α-subunit)

The proteasome is a multicatalytic proteinase complex with a highly ordered ring-shaped core structure. The core structure is composed of 4 rings of 28 non-identical subunits; 2 rings are composed of 7 α- subunits and 2 rings are composed of 7 β subunits. Proteasomes are distributed throughout eukaryotic cells at a high concentration and cleave peptides in an ATP/ubiquitin-dependent process in a non-lysosomal pathway (taken from NCBI Gene). Note that 6 of 7 PSMAs are found at k6 and the missing PSMA5 is found at k4. (see fig. 32 and table 22)

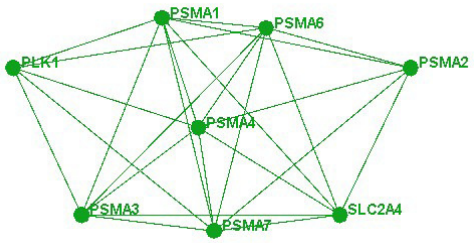


Figure 31: Six of seven PMAs of the proteasome α-subunit (k6 c5).

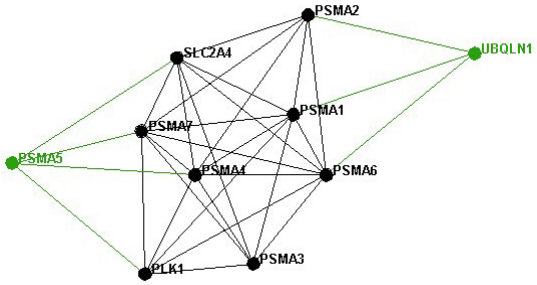


Figure 32: All seven proteins from the proteasome α-subunit at k4 c119.

Along with PSMAs three other proteins are included. SLC2A4 is the Glucose-transporter 4 and the top-hub of the network with 770 interactions. PLK1 (Polo-like kinase 1) is shown to be an important mitotic regulator of proteasome activity ([Feng et al., 2001]) although very little literature is available about the interrelationship with PLK1 and the proteasome. PLK1 is also linked with the PSMBs (β-subunit of the proteasome) using the graph density method (not shown).

Table 22: The Proteasome (α-subunit).

k	c	V	E	CC_{n}	Biological Activity	Reference
6	5	8	25	0,86	Protein Degradation	NCBI Gene
5	39	8	25	0,86		
4	119	10	32	0,64		

Replication Factor C

The elongation of primed DNA templates by DNA polymerase delta and DNA polymerase epsilon requires the accessory proteins proliferating cell nuclear antigen (PCNA) and replication factor C (RFC). RFC is a protein complex consisting of five distinct subunits. The core complex possesses DNA-dependent ATPase activity, which was found to be stimulated by PCNA in an in vitro system. (NCBI Gene). Figure 33 shows PCNA highly interwired with the RFC2-4 in the core.

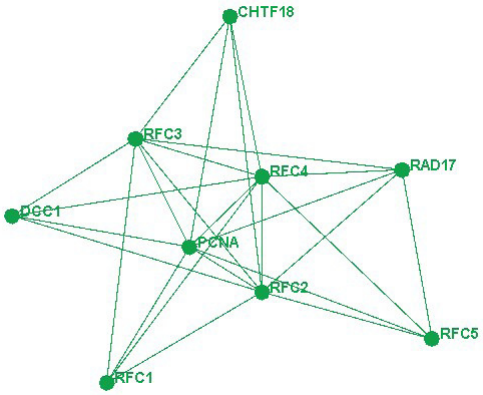


Figure 33: Proteins of the Replication Factor C (k5 c1).

Table 23: Replication Factor C.

k	c	V	E	CC_{n}	Biological Activity	Reference
5	1	9	26	0,64	Replication Factor	NCBI Gene

TCP1 ring complex (TRiC)

CCT6A encodes a molecular chaperone that is a member of the chaperonin containing TCP1 complex (CCT), also known as the TCP1 ring complex (TRiC). This complex consists of two identical stacked rings, each containing eight different proteins. Unfolded polypeptides enter the central cavity of the complex and are folded in an ATP-dependent manner. The complex folds various proteins, including actin and tubulin (NCBI Gene). Five proteins are found here (see fig. 34 and table 24).

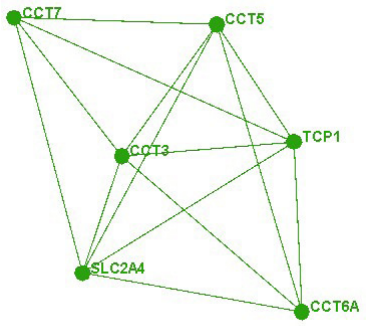


Figure 34: TCP1 ring complex important for protein folding (k5 c12).

Table 24: The TCP1 ring complex.

k	c	V	E	CC_{n}	Biological Activity	Reference
5	12	6	14	0,9	TCP1 ring complex / Folding	NCBI Gene

Origin Recognition Complex

The origin recognition complex (ORC) is a highly conserved six subunits protein complex essential for the initiation of the DNA replication in eukaryotic cells. Studies in yeast demonstrated that ORC binds specifically to origins of replication and serves as a platform for the assembly of additional initiation factors such as CDC6 and MCM proteins. ORC2L forms a core complex with ORC3L, -4L, and -5L. It also interacts with CDC45L and MCM10, which are proteins known to be important for the initiation of DNA replication (NCBI). All six subunits ORC1L-6L as well as the important interactors CDC45L and MCM10 are reconstituted in figure 36. CDC6 and MCM proteins are also shown for k6. (See fig. 35)

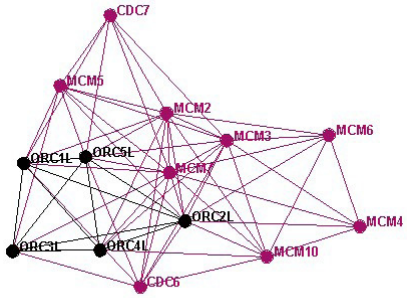


Figure 35: The Origin Recognition Complex showing five of six known ORC-subunits (k6 c10) as well as CDC6 and MCM proteins.

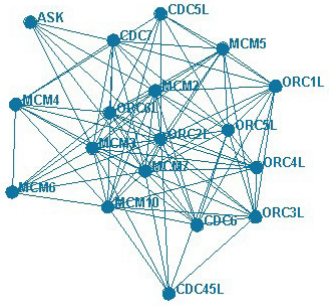


Figure 36: The Origin Recognition Complex showing all six known ORC-subunits (k5 c24).

Table 25: The Origin Recognition Complex.

k	c	V	E	CC_{n}	Biological Activity	Reference
6	10	14	61	0,62	Origin Recognition Complex	NCBI Gene
5	24	18	92	0,55		

Further *Communities*

Further *communities* are manually annotated (not shown) to apoptosis-associated proteins, BCL and Bak, Activin Receptor Complex , Integrin complexes, Tumor Necrosis Factor associated proteins, period (circadian pattern), liprin (axon guidance), synaptic vesicle transport assembly, fusion of snaptic vesicles (VAMPS/

Snap25), glutamate receptor complex, DNA methyltransferase activity, sumoylation, growth arrest and DNA-damage, general transcription factors, cAMP response element activity (Creb), bone morphogenic proteins (BMPs), neurotransmitter secretion, rab-proteins, interleukin-1 receptor-associated kinases (IRAKs), eukaryotic translation initiation factor 3 complex, sodium channel, coagulation factor 2, collagen chains, sorting nexin, transcription elongation factor, poliovirus receptor-related 1 complex, silencing, calcium-dependent membrane-binding or MAP Kinases (data not shown).

Table 26 shows the proteins that have a high abundance in different *communities*, the top hubs SLC2A4 and TP53 participate in many *communities* despite their low clustering properties. This finding might represent that the two proteins function as date-hubs (hubs mediating diverse functions).

Table 26: Top 10 proteins having high *community* memberships.

Protein	# of *communities*	*degree*	CC1(v)
SLC2A4	62	770	0,0017
TP53	29	261	0,0207
HDAC1	20	130	0,0515
CTNNB1	17	133	0,0355
CREBBP	16	202	0,0321
SRC	16	200	0,0407
GRB2	15	198	0,0449
PCNA	15	81	0,0398
EGFR	14	147	0,0494
ESR1	14	171	0,0477

Communities important for immunobiology

NF-κB complex

NF-κB is a transcription regulator that is activated by various intra- and extracellular stimuli such as cytokines, oxidant-free radicals, ultraviolet irradiation, and bacterial or viral products. Activated NF-κB translocates into the nucleus and stimulates the expression of genes involved in a wide variety of biological functions. Inappropriate activation of NF-κB has been associated with a number of inflammatory diseases while persistent inhibition of NF-κB leads to inappropriate immune cell development or delayed cell growth (NCBI Gene). (See fig. 37 and table 27)

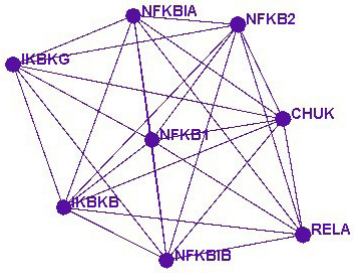


Figure 37: Proteins of the NF-κB complex (k7 c12).

Table 27: The NF-κB complex.

k	c	V	E	CC_{n}	Biological Activity	Reference
7	12	8	27	0,95	NF-κB complex	NCBI Gene

Transforming Growth Factor-β

Transforming Growth Factor-β (TGF-β) is a multifunctional peptide that controls proliferation, differentiation, and other functions in many cell types. TGF-β acts synergistically with TGF-α in inducing transformation. It also acts as a negative autocrine growth factor. Dysregulation of TGF-β activation and signaling may result in apoptosis. Many cells synthesize TGF-β and almost all of them have specific receptors for this peptide. TGF-β1, TGF-β2, and TGF-β3 all function through the same receptor signaling systems (NCBI Gene). (See fig. 38 and table 28)

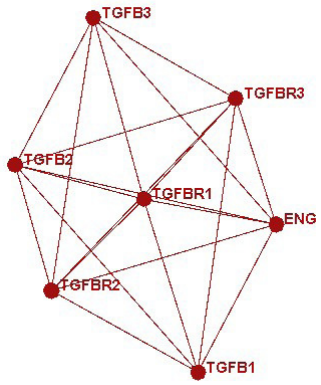


Figure 38: Proteins of Transforming Growth Factor-β (k6 c16).

Table 28: Transforming Growth Factor-β.

k	c	V	E	CC_{n}	Biological Activity	Reference
6	16	7	20	0,93	Transforming Growth Factor-β	NCBI Gene

B-cell antigen receptor complex

Lymphocytes proliferate and differentiate in response to various concentrations of different antigens. The ability of the B cell to respond in a specific, yet sensitive manner to the various antigens is achieved with the use of low-affinity antigen receptors. CD19 encodes a cell surface molecule which assembles with the antigen receptor of B lymphocytes in order to decrease the threshold for antigen receptor-dependent stimulation. The B lymphocyte antigen receptor is a multimeric complex that includes the antigen-specific component, surface immunoglobulin (Ig). Surface Ig non-covalently associates with two other proteins, Ig-α and Ig-β, which are necessary for expression and function of the B-cell antigen receptor. CD79B encodes the Ig-β protein of the B-cell antigen component. (taken and adapted from NCBI Gene). (See fig. 39 and table 29)

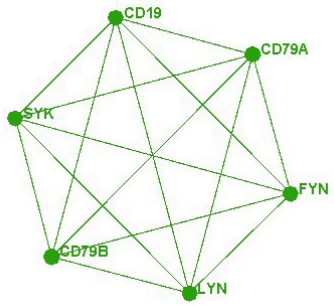


Figure 39: Proteins of the B-cell antigen receptor complex (k6 c20).

Table 29: The B-cell antigen receptor complex.

k	c	V	E	CC_{n}	Biological Activity	Reference
6	20	6	15	1	B-cell antigen receptor complex	NCBI Gene

Transmembrane 4 superfamily, tetraspanin family

CD63 (TSPAN30), CD81 (TSPAN28), CD9 (TSPAN29) and CD151 (TSPAN24) are members of the transmembrane 4 superfamily, also known as the tetraspanin family. Most of these members are cell-surface proteins that are characterized by the presence of four hydrophobic domains. The proteins mediate signal transduction events that play a role in the regulation of cell development, activation, growth and motility. CD63 is a cell surface glycoprotein that is known to complex with *integrins*. It may function as a blood platelet activation marker. Deficiency of this protein is associated with Hermansky-Pudlak syndrome. Also this gene has been associated with tumor progression. (See fig. 40 and table 30)

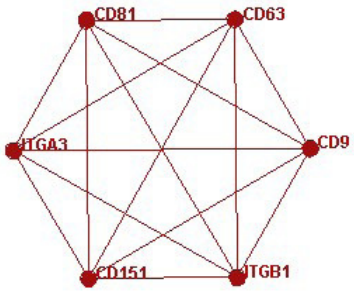


Figure 40: Proteins of the Tetraspanin/ transmembrane 4 superfamily (k6 c23).

EWI (PTGFRN) proteins, through their direct interaction with ezrin-radixin-moesin (ERM) proteins, act as *linkers* to connect *tetraspanin*-associated microdomains to actin cytoskeleton regulating cell motility and polarity [Sala-Valdés et al., 2006]. Using mAb C9BB as a tool, Yang et al. [Yang et al., 2006] showed that cell surface CD9 homoclustering is promoted by expression of $\alpha 3\beta 1$ (ITGA3) and $\alpha 6\beta 4$ (ITGB4) integrins and by palmitoylation of the CD9 and $\beta 4$ proteins. In addition to ITGA3, ITGA6 and ITGB1 are also associated with this complex (fig. 41 and table 30).

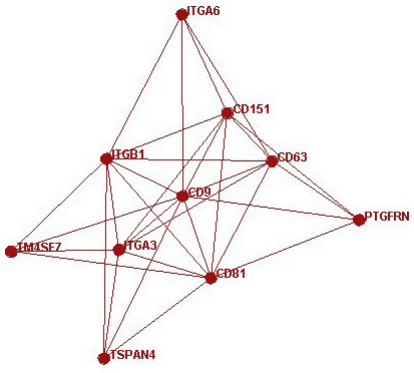


Figure 41: Tetraspanin/ transmembrane 4 superfamily in complex with integrins

Table 30: The Tetraspanin/ transmembrane 4 superfamily.

k	c	V	E	CC_{n}	Biological Activity	Reference
6	23	6	15	1	the Tetraspanin/ transmembrane 4 superfamily	[Sala-Valdés et al., 2006]
5	18	10	1	0,61		

Major histocompatibility complex, class II, DR α

HLA-DRA is one of the HLA class II α chain paralogues. This class II molecule is a heterodimer consisting of an α and a β chain, both anchored in the membrane. It plays a central role in the immune system by presenting peptides derived from extracellular proteins. Class II molecules are expressed in antigen presenting cells (APC: B lymphocytes, dendritic cells, macrophages). DRA does not have polymorphisms in the peptide binding part and acts as the sole α chain for DRB1, DRB3, DRB4 and DRB5 (NCBI Gene). (See fig. 42 and table 31)

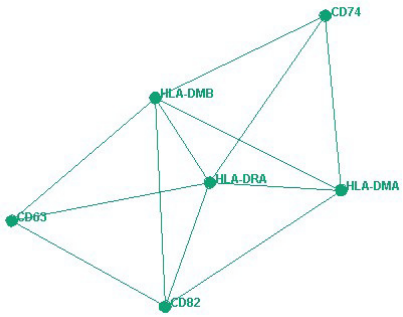


Figure 42: Proteins of the major histocompatibility complex, class II, DR α (k4 c121).

Table 31: The major histocompatibility complex, class II, DR α.

k	c	V	E	CC_{n}	Biological Activity	Reference
4	121	6	12	0,7	major histocompatibility complex, class II, DR α	NCBI Gene

B-lymphocyte activation antigen CD80

The B-lymphocyte activation antigen B7-1 (CD80, B7) provides regulatory signals for T lymphocytes as a consequence of binding to the CD28 and CTLA4 ligands of T cells (NCBI Gene). (See fig. 43 and table 5.4.3)

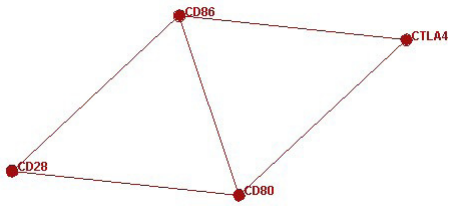


Figure 43: Proteins of the B-lymphocyte activation antigen CD80 (k3 c106).

The B-lymphocyte activation antigen CD80.

k	c	V	E	CC_{n}	Biological Activity	Reference
3	106	4	5	0,67	B-lymphocyte activation antigen CD80	NCBI Gene