

## Insights into biological information processing: structural and dynamical analysis of a human protein signalling network

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2008 J. Phys. A: Math. Theor. 41 224013

(<http://iopscience.iop.org/1751-8121/41/22/224013>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 171.66.16.149

The article was downloaded on 03/06/2010 at 06:52

Please note that [terms and conditions apply](#).

# Insights into biological information processing: structural and dynamical analysis of a human protein signalling network

**Alberto de la Fuente, Giorgio Fotia, Fabio Maggio, Gianmaria Mancosu and Enrico Pieroni**

CRS4 Bioinformatica, Parco Tecnologico POLARIS, Ed.1, Loc Piscinamanna, Pula, Italy

E-mail: [alf@crs4.it](mailto:alf@crs4.it)

Received 20 September 2007, in final form 5 November 2007

Published 21 May 2008

Online at [stacks.iop.org/JPhysA/41/224013](http://stacks.iop.org/JPhysA/41/224013)

## Abstract

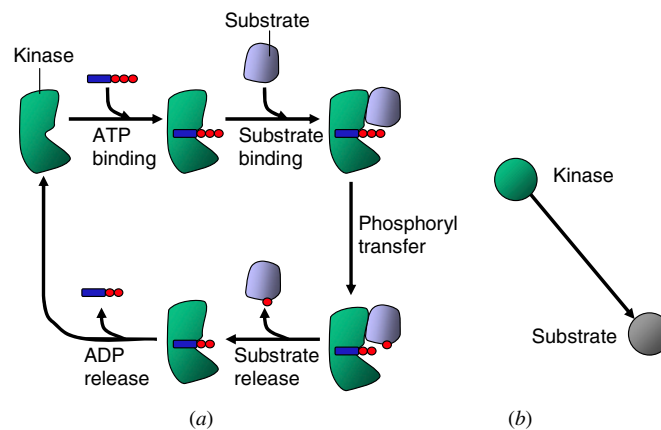
We present an investigation on the structural and dynamical properties of a 'human protein signalling network' (HPSN). This biological network is composed of nodes that correspond to proteins and directed edges that represent signal flows. In order to gain insight into the organization of cell information processing this network is analysed taking into account explicitly the edge directions. We explore the topological properties of the HPSN at the global and the local scale, further applying the generating function formalism to provide a suitable comparative model. The relationship between the node degrees and the distribution of signals through the network is characterized using degree correlation profiles. Finally, we analyse the dynamical properties of small sub-graphs showing high correlation between their occurrence and dynamic stability.

PACS numbers: 89.75.Hc, 87.10.+c, 87.14.Ee, 89.70.+c

(Some figures in this article are in colour only in the electronic version)

## 1. Introduction

Living cells are continuously subjected to many stimuli from their environment, which require appropriate responses. As for maintaining stable their internal condition or performing cellular functions, cells are supplied with dedicated biochemical machinery of which proteins are the key molecular entities. While this machinery is extensively characterized at the molecular detail, a big effort is made by the system biology community to improve knowledge about its large scale organization. For this purpose biochemical machinery can be successfully represented as networks, where the nodes are proteins and the edges indicate an interaction



**Figure 1.** (a) Example of a catalytic cycle of a substrate (protein) phosphorylation by a kinase (protein) [6]. The adenosine triphosphate (ATP) and the substrate bind to the active site of the kinase. One phosphate group is transferred to a particular site of the substrate, which subsequently is released. Upon release of the remaining adenosine diphosphate (ADP), the kinase returns to the initial state. (b) Abstraction of this process in the PSN.

between adjacent proteins. Investigating the structural and dynamical properties of such networks will provide insight in the organization of biological processes.

Protein interaction networks (PINs) are undirected networks in which edges between nodes indicate physical binding of proteins and have been subjected extensively to network analysis [1–4]. Since no flow of material or information occurs, analysing PINs gives no information on causality or cascades of chemical events. A more interesting network is obtained by looking at the activity of proteins as information signal processing. For example, information signals at the surface of cells (e.g. hormones or nutrients) trigger cascades of chemical reactions, which directly results in altered activities of proteins, and eventually leads to changes in biochemical processes, such as activation of gene expression, alternation of metabolism, etc. We define these networks as protein signalling networks (PSNs) according to [4]. In PSNs, the nodes correspond to levels of chemically modified states of proteins and directed edges to signal flows established through chemical modifications. In such a network, source proteins pass their signals to the target proteins by chemically modifying them. It is important to explicitly point out that, in contrast to PINs, PSNs are directed and correspond to actual dynamical regulatory systems in which information is flowing over the network. Analysis of the directed structure of PSNs will provide better understanding of the nature of information processing by cells.

Here, we present results on structural and dynamic analysis of a human PSN (HPSN) proposed by Linding *et al* [5] and publicly available from <http://networkin.info>. This network is based on a post-translational protein modification known as ‘phosphorylation’, in which a phosphate group is transferred to a protein (substrate) by means of specialized proteins called kinases (figure 1(a)). This catalytic process can be rendered as a directed edge between a kinase (source) and a substrate (target), to indicate a clear direction in the signal flows (figure 1(b)). A phosphorylated protein changes its activation state and passes signals to other proteins or regulates, for example, gene expression or metabolic processes. Being proteins, kinases can be phosphorylated as well, giving origin to chains of phosphorylation events.

In the following, we start by characterizing the network in terms of its global features. In section 3, we identify topological components and apply the generating function formalism to

emphasize the difference between the HPSN and similar networks. In section 4, we investigate the node–degree correlations to gain some insight into the way signals are propagated through the network. In an effort to relate structure to dynamics, in section 5 we study the stability properties of sub-graphs in the network. In section 6, we summarize our findings and draw some conclusions.

## 2. Basic features and topological characterization of the HPSN

The HPSN has 5189 edges ( $e$ ) and 1810 nodes ( $n$ ), hence it is quite sparse with a relative edge density of  $e/[n(n - 1)] = 0.00159$ . It contains 69 source nodes, 20 of which do not receive any input and has 1790 targets, 1741 of which do not have any outputs. Only 49 nodes are both sources and targets. The maximum number of targets of a single node is 533 and the maximum number of inputs into a single node is 19. In the HPSN there are 11 cycles of two edges (mutually regulating nodes) and the in-degree ( $k_{\text{in}}$ ) and out-degree ( $k_{\text{out}}$ ) have an average value of 2.87.

The in-degree distribution is well fit by an exponential law:  $P(k_{\text{in}}) \propto \exp(-\alpha k_{\text{in}})$  with  $\alpha = 0.396$  ( $R^2 = 0.9728$ ). A similar trend has been observed for the in-degree distribution of transcriptional regulatory networks [7] and gene networks [8]. Such a law may be the result of a growing mechanism without preferential attachment [9] or due to a growing mechanism with preferential attachment, but with the constraint that each node has a limited capacity to receive inputs [10]. The latter mechanism seems plausible for information processing networks: too many information signals into a node will merely ‘confuse’ it. Since there are only 69 source nodes in this network, nothing much can be said about the precise law underlying the out-degree distribution. What can be said, however, is that there are few ‘hubs’ and a much larger fraction of ‘non-hubs’. The joint in–out degree distribution  $P(k_{\text{in}}, k_{\text{out}})$  has very small dependency (Spearman rank correlation coefficient:  $-0.0682$ , Pearson correlation:  $-0.0108$ ).

The shortest path-length distribution is a slightly right-skewed bell-shaped function with average value of 3.437 (very small compared to the number of nodes: typical property of a small world [11]) and the network diameter is 8. These features indicate the presence of very few intermediate layers. There are 28613 pairs of nodes where a path exists only in one direction (1.75% of all possible pairs) and 301 mutually linked pairs, indicating that each node can reach only a tiny fraction of other nodes. This gives an immediate indication that information is confined to sub-sets of the network. Studying the degree mixing (see section 4) will shed more light on how information is confined.

The upstream cluster coefficient defined for a given node  $i$  is the number of triangles ( $N_{\text{tr}}$ ) the node forms with two input neighbours divided by the number of all possible such triangles:  $C_i^{\text{ups}} = N_{\text{tr}}/k_{\text{in}}(k_{\text{in}} - 1)$  [7]. It measures how likely a node is part of an upstream clique. The average upstream clustering coefficient is 0.117, showing that on average a node can have only about 12% of the incoming neighbours that are adjacent. The upstream clustering coefficient distribution fits quite well a decay behaviour  $C^{\text{ups}}(k_{\text{in}}) = 0.6326 k_{\text{in}}^{-0.73}$  ( $R^2 = 0.8294$ ), that is a typical signature of an underlying hierarchical structure [12, 13]. The downstream clustering coefficient distribution roughly shows decay behaviour, though a good fit to a power law cannot be obtained. The value of the average downstream clustering coefficient is 0.00386, two orders of magnitude smaller than its upstream counterpart. The ratio of the two average cluster coefficients indicates that it is much more likely to have a clique in the input rather than in the output neighbourhood. Next, we compare these results to the mean properties of the ensemble of networks with the same degree distribution by using the generating function (GF) formalism [14]. Differences between the results from this analytic framework and

the observations in the HPSN may indicate the action of evolutionary forces having shaped the network for functional purposes. Given the joint distribution  $P(k_{\text{in}}, k_{\text{out}})$ , the generating function is defined as

$$G(x, y) = \sum_{k_{\text{in}}, k_{\text{out}}} P(k_{\text{in}}, k_{\text{out}}) x^{k_{\text{in}}} y^{k_{\text{out}}}.$$

The upstream and downstream clustering coefficients can be then derived as [7]

$$N_{\text{tr}} = \frac{\langle k_{\text{in}}(k_{\text{in}} - 1) \rangle}{\langle k_{\text{in}} \rangle} \frac{\langle k_{\text{out}}(k_{\text{out}} - 1) \rangle}{\langle k_{\text{out}} \rangle} \frac{\langle k_{\text{in}} k_{\text{out}} \rangle}{\langle k_{\text{in}} \rangle} = \frac{\partial^2 G}{\partial x^2} \frac{\partial^2 G}{\partial y^2} \frac{\partial^2 G}{\partial x \partial y} \left( \frac{\partial G}{\partial x} \right)^{-3} \Bigg|_{x=y=1}$$

$$N_{\text{in}} = \langle k_{\text{in}}(k_{\text{in}} - 1) \rangle = \frac{\partial^2 G}{\partial x^2} \Bigg|_{x=y=1}; \quad N_{\text{out}} = \langle k_{\text{out}}(k_{\text{out}} - 1) \rangle = \frac{\partial^2 G}{\partial y^2} \Bigg|_{x=y=1}$$

$$C_i^{\text{ups}} = N_{\text{tr}}/N_{\text{in}}; \quad C_i^{\text{dns}} = N_{\text{tr}}/N_{\text{out}}.$$

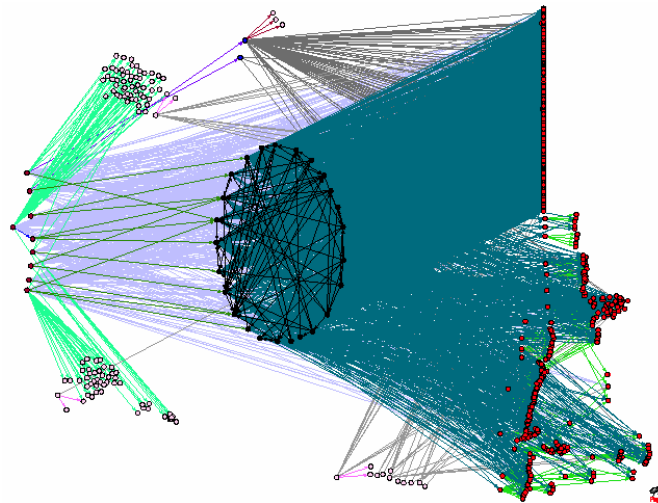
The upstream clustering coefficient estimate (0.124) roughly agrees with that observed (0.117), while the predicted value of downstream clustering coefficient (0.0019) shows a relatively larger difference compared to that observed (0.00386). Finally, taking into account the predicted relative size  $S$  of the giant component (see section 3), the GF formalism allows us to compute the average shortest path [14]:  $l = \log(Sn/z_1)[\log(z_2/z_1)]^{-1} + 1$ . The obtained value is 3.081, a bit less than observed in HPSN (3.437).

### 3. HPSN components and comparison with analytical predictions using generating function framework

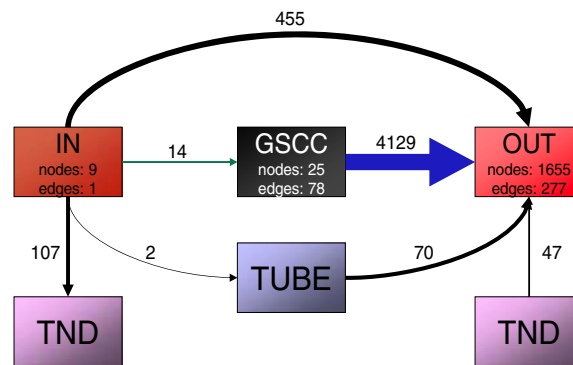
#### 3.1. Global component analysis

The global topology of many directed graphs, such as the WWW [15], metabolic networks [16] and gene networks [8] resembles a ‘bow-tie’. As shown in figure 2, the HPSN topology fits such a structure as well.

A strongly connected component (SCC) is defined as a maximal set of nodes in which for each pair of nodes  $a$  and  $b$  there exists a directed path from  $a$  to  $b$  and vice versa. The giant strongly connected component (GSCC) is the largest of the SCCs of the network. The IN component nodes can reach the GSCC nodes through a directed path, but not vice versa. The OUT component nodes can be reached from the GSCC but not vice versa. The TUBE contains nodes connecting IN to OUT. All nodes that do not belong to previous components are grouped in the TENDRIL. In the analysed HPSN the GSCC consists of 25 nodes ( $\sim 1.4\%$ ) and 78 edges. There are nine nodes into the IN ( $\sim 0.5\%$ ), 1655 into the OUT ( $\sim 91\%$ ), two into the TUBE and 119 into the TENDRIL. Comparing the relative sizes of the components in the HPSN with other networks shows different results. For example, in the WWW [15] the nodes are almost equally distributed over the components (GSCC  $\sim 28\%$ , IN  $\sim 21\%$  and OUT  $\sim 21\%$ ). In the English Wikipedia in June 2004, the IN and OUT components are of about the same size ( $\sim 7\%$ ) with a much larger GSCC of  $\sim 82\%$  [18]. On the other hand, several biological networks show the proportions observed in HPSN. For example, the *E. coli* metabolic network features 15% for the IN and 30% for the OUT, with a GSCC of 40% [19]. A yeast gene network (YGN) [8] indeed is very similar to the HPSN in terms of the component sizes: IN  $\sim 2\%$ , OUT  $\sim 73\%$  and GSCC  $\sim 14\%$ . This observation could be explained by the fact that both the HPSN and YGN are regulatory networks where the number of regulators is small compared to the number of targets.



**Figure 2.** Bow-tie structure of HPSN. The picture was obtained by applying several layout algorithms in Pajek [17]. The GSCC in the HPSN resembles a ‘brain’ in which the signals coming from the IN component are processed and subsequently transmitted to the OUT component.



**Figure 3.** Compact representation of the bow-tie structure of HPSN. Values along arrows indicate how many edges join the components. Inside IN, OUT and GSCC boxes, number of nodes and internal edges are reported. The three edges between TUBE and TENDRIL (TND) are omitted.

The bow-tie is a purely node-oriented classification. Focusing on edges, we discovered the presence of IN nodes ‘talking’ directly to OUT nodes (thus without passing through GSCC or TUBE). These shortcuts establish fast transmission of signals by circumventing complex processing by the GSCC, confirming what has been found in the previous section about small world feature. Figure 3 reports the number of edges inside and between each component: the largest part of the information (counted as the number of edges) flows between IN and OUT through the GSCC (14 and 4129 edges, respectively), whereas there are 455 shortcut edges from IN to OUT.

### 3.2. Topological characterization of GSCC

Since the GSCC is the most interesting component, we applied the previous topological analysis (see section 2) specifically to it. The GSCC has an edge density of 0.13, two order of magnitude higher than the whole HPSN (0.00159). The in-degree distribution is still roughly exponential with a coefficient  $\alpha = 0.24$  ( $R^2 = 0.828$ ), resulting in a fatter tail. With respect to the HPSN, the average in and out degree increases to 3.12, the average path length reduces to 2.93 and the diameter to 6. The upstream clustering coefficient is 0.234 and the downstream clustering coefficient is 0.227. These values are closer to each other and both higher than the HPSN counterparts, as expected since GSCC is denser. Both upstream and downstream clustering coefficient distribution have decay behaviour roughly mimicking the  $1/\text{degree}$  behaviour. By using the generating functions we predict the values of 0.149 and 0.119 for the average downstream and upstream clustering coefficient, respectively, which are quite different from the observed values. Finally, the predicted average path length for GSCC is 2.513.

A threshold for the GSCC appearance is identified by  $\sum_{k_{\text{in}}, k_{\text{out}}} (2k_{\text{in}}k_{\text{out}} - k_{\text{in}} - k_{\text{out}})P(k_{\text{in}}, k_{\text{out}}) \geq 0$ . For the HPSN this sum is 9.434, showing that the giant component transition already took place. This allows us to use the GF approach to predict the size of the bow-tie components [14, 20]. We define the generating functions separately for the in- and out-degree distribution of a node:  $F_0(x) = G(x, 1)$ ,  $G_0(y) = G(1, y)$ . We also need to define the distribution of degrees for a randomly chosen neighbour of a given node:  $F_1(x) = \langle k_{\text{in}} \rangle^{-1} \partial G / \partial y(x, 1)$ ,  $G_1(y) = \langle k_{\text{out}} \rangle^{-1} \partial G / \partial x(1, y)$ . Sizes of components can be defined as follows:  $S + O = 1 - F_0(u)$ ,  $S + I = 1 - G_0(v)$ ,  $S = 1 - F_0(u) - G_0(v) + G(u, v)$ , where  $u$  and  $v$  are solutions of  $u = F_1(u)$ ,  $v = G_1(v)$  and  $I$ ,  $O$ ,  $S$  are the size of the IN, OUT and SCC components, respectively. The predicted relative sizes are the following: SCC + IN components 1.62% (observed 1.88%), GSCC + OUT 92.48% (observed 92.82%). As for the relative size of the GSCC itself we found 1.26% against the observed value of 1.38%.

## 4. Correlation profiles analysis of the HPSN

Further insight in the topology of the HPSN can be obtained using degree correlation profiles analysis. Correlation profiles have been introduced in [21, 22] to characterize high-level topological properties of complex networks. Following this approach, the correlation in degrees of adjacent nodes is compared between the real network and a properly randomized counterpart that preserves some of its low-level topological properties. Implicitly, computation of correlation between degrees of adjacent nodes of a directed network results in a classification of its edges in terms of out- and in-degree of source and target nodes. In fact, we can envisage four different classification schemes (figure 4), denoted as out-in, in-in, in-out, out-out, depending if in- or out-degree is considered, respectively, of the upward node or of the downward node of a given edge. Actually, we extended the approach proposed in [21, 22] by considering further edge classification schemes in addition to the out-in used by authors. The proposed four schemes, in turn, may be associated with specific functional patterns.

Since we are interested in categorizing in- and out-degree in four classes, zero, low, intermediate and high degree, we define small in- (out-) hubs, and large in- (out-) hubs, depending if either  $7 \leq k_{\text{in}} < 14$  ( $200 \leq k_{\text{out}} < 400$ ) or  $14 \leq k_{\text{in}} < 20$  ( $400 \leq k_{\text{out}} < 600$ ). The remaining nodes, characterized by either low nonzero in- or out-degree, that is  $1 \leq k_{\text{in}} < 7$  and  $1 \leq k_{\text{out}} < 200$ , are defined as non-hubs. In-hubs can be seen as signal integrators; combining many inputs to form a complex output to other nodes or to establish the output response.



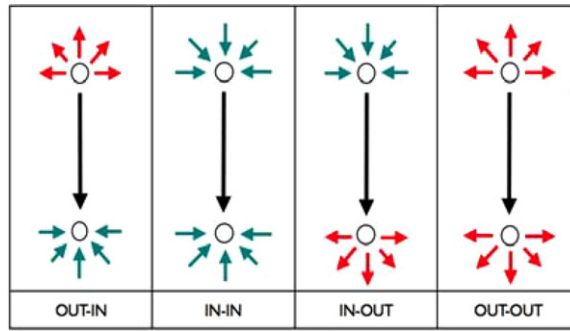


Figure 4. Classification schemes used in correlation profile analysis.

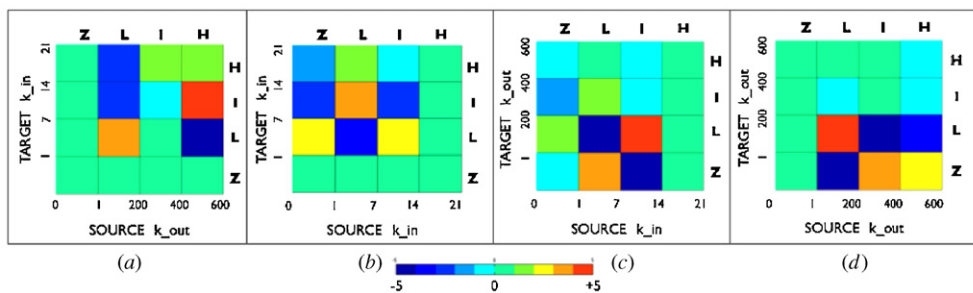


Figure 5. Z-score of node-degree-correlations. Connectivities were binned according to the maximum in- or out-degree in four different classes, zero (Z), low (L), intermediate (I) and high (H). (a) Z-score for out-in correlation profile. (b) Z-score for in-in correlation profile. (c) Z-score for in-out correlation profile (d) Z-score for out-out correlation profile.

Out-hubs can be seen as signal distributors; a signal originating or coming through such nodes reaches many other nodes. Comparison of the observed properties with an appropriate null model may also reveal if some of these organizational patterns assume a particular relevance, as a result of a significant deviation from chance. Following [21], the null model may be obtained by randomly rewiring the edges of the HPSN, using a local rewiring algorithm based on edge swapping. In this null model, all nodes conserve exactly the same in- and out-degree of the original network, whereas the edges are randomly reshuffled, resulting in a randomization of any correlation between degree values of connected nodes. This random counterpart is used to identify those topological patterns which are significantly over- or under-represented in the real network. Deviations from null model are evaluated in terms of Z-score, defined as  $Z(k_i^s, k_i^t) = [N(k_i^s, k_i^t) - \langle N_r(k_i^s, k_i^t) \rangle] [\Delta_r(k_i^s, k_i^t)]^{-1}$ , where  $\Delta_r$  is the standard deviation of  $N_r$  in 1000 realizations of the randomized network.  $N$  is the number of times that a directed edge is observed in the real network between a node with source node degree  $k_i^s$  and target node degree  $k_i^t$  ( $i = \text{in}, \text{out}$ );  $\langle N_r \rangle$  is the average number of occurrences of such edges in the randomized version of the network. Large positive (negative) values of Z-score indicate that edges in the HPSN network are over- or under-represented compared to the null model.

Figure 5(a) displays the Z-score of node-degree correlations of the *out-in* classification scheme, showing that in the present network high-degree source nodes (large out-hubs, i.e. signal distributors) preferentially regulate intermediate degree target nodes (small in-hubs).



If we define peers as those elements that belong to the same class of in- or out-degree (e.g., low, intermediate or high), we may conclude that in the present case, information does not flow significantly between peers. Finally, the information transfer is under represented between non-hubs and small and large in-hubs. As a result, information flows preferentially along a hierarchical path. This suggests the existence of relatively semi-independent modules hierarchically organized, which is a feature already observed in many cellular processes, and in molecular networks in particular [21, 23].

Figure 5(b) shows the Z-score of connectivity correlation of the *in-in* classification scheme. Here, we investigate which is the flow of information between signal integrators of different importance, and the information flow from zero in-degree nodes to small or large in-hubs, i.e. signal integrators. Specifically, it is found that small in-hubs preferentially output to non-hubs, and non-hubs preferentially output to small in-hubs. Signal integrators, therefore, appear not to exchange information between signal integrators of the same importance.

Figure 5(c) shows the Z-score of connectivity correlation of the *in-out* classification scheme. Here, we investigate which is the flow of information between signal integrators and signal distributors and the information flow from signal integrators to zero in-degree nodes. It is important to note that for the present classification scheme, we focus our attention on target zero out-degree nodes, since it is found that more than 80% of network edges are characterized by the property of having zero out-degree target nodes. Edge enrichment is found for connection from low in-degree nodes to zero out-degree nodes, whereas edge suppression is found for connections from intermediate in-degree nodes to zero out-degree nodes. A preferential path of information processing from non-hubs to zero out-degree nodes is observed. Information flow from small in-hubs to zero out-degree nodes is instead greatly suppressed. Since connection of in-hubs to out-hubs appears not to be largely over-represented, it may be conjectured that the topological structure of the present network does not favour distribution of complex signals over the whole network.

Finally, figure 5(d) shows the Z-score of connectivity correlation of the *out-out* classification scheme. Here, we investigate which is the flow of information between signal distributors of different importance and the information flow from signal propagators to zero out-degree nodes. Also in the present case, we focus our attention on source zero out-degree nodes. Inspection of the figure shows that large and small out-hubs communicate preferentially with zero out-degree out nodes, which are the end points of the signal flow. Connections between non-hubs are over-represented, while from non-hubs to zero out-degree nodes are under-represented. This indicates that the topology of the network does not favour fast propagation of signals over the whole network, which would be the case if out-hubs would output to other out-hubs. It also shows that most of the signals that are distributed over the network are carried through low out-degree nodes.

## 5. Sub-graphs and stability

### 5.1. Sub-graph analysis

The objects of this section are 3- and 4-node sub-graphs. Sub-graphs occurring at a frequency that is significantly different than expected by chance (usually called motifs) may be seen as elementary building blocks of complex networks [24, 25]. For HPSN, sub-graphs were identified by running the software FanMod [26]. By comparing the frequency of occurrence of each sub-graph in the HPSN with respect to a set of randomized networks (1000 in the present study) we identified several motifs.

**Table 1.** Three- and four-node sub-graphs mentioned in the text, along with their Z-score.















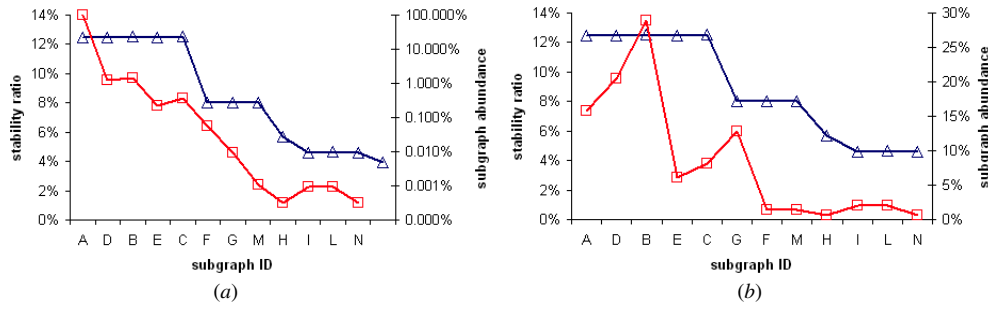
													
A	B	C	D	E	F	G	H	I	L	M	N	O	P
0.21	2.71	-5.96	1.34	-2.77	5.83	-1.04	-1.43	2.33	1.14	0.35	1.47	16.69	5.51

Table 1 displays the 3- and 4-node sub-graphs mentioned in this work, along with their Z-score. The feed-forward loop (FFL, marked as E) that is often observed to be over-represented in biological information processing networks, such as transcriptional regulatory networks and neuronal networks [24], is actually under-represented in the HPSN ( $Z = -2.77$ ). However, sub-graph F is highly over-represented and can be seen as a superposition of two FFLs. It seems that this information processing system required a slightly advanced form of the simple FFL. This sub-graph has also been observed to be over-represented in other signal transduction networks [24, 25]. Sub-graph F is composed by a pair of nodes with reciprocal regulation, co-regulating the same node. It can then be interpreted as a backup or a concerted co-regulation system. In the former case, one of the two source nodes can act as a backup for the other source node when it fails, while in the latter the source pair is deciding together what information pass and how to the target node. In common with food webs, cascade sub-graph B is over-represented in the HPSN. The feed-back loop (sub-graph I) only occurs in six instances in the network, but still is significantly over-represented in common with electronic circuits (digital fractional multipliers) [24]. Among 4-node sub-graphs, the bi-fan (sub-graph O) in which two nodes jointly input in two other nodes is strongly over-represented. Interestingly, this indeed is in common with other biological information processing networks and electronic circuits. Also, an advanced form of the bi-fan, sub-graph P, in which the two input nodes also communicate with one another, is strongly over-represented, with similar interpretations as for F.

## 5.2. Stability and sub-graphs

Analysis of dynamics provides information on the response of biological systems to external perturbations. Systems that upon a perturbation exhibit a damped response pointing to the initial steady state are classified as stable. A network can be represented by the Jacobian matrix  $\mathbf{J}$ , essentially a weight matrix, corresponding to a linear set of ordinary differential equations (see, e.g., [25, 27]). The network is stable if all the eigenvalues of  $\mathbf{J}$  have negative real part. If in addition the corresponding imaginary parts are zero, stable solutions are monotonic functions of time (no damped oscillations).

Stability of a network depends on its structure and on the weights of edges between nodes. The goal of the study here is to quantify the effect of the network structure rather than the weights. To average out the effect of the actual weights we study the stability over a range of different weight values. We measure the probability for a given network to be stable resulting from a set of trial configurations, for which entries of  $\mathbf{J}$  are drawn from a given distribution. We here specifically investigate the dynamic stability, but note that this approach can be used to quantify probabilities for other dynamic properties as well, such as the ratio of oscillatory versus non-oscillatory realizations. In addition, this approach allows detecting certain types of bifurcations in the specified range of parameters [27]. Similarly to the analysis performed by Prill *et al* [25], we check the relation between abundance of the 12 occurring 3-node sub-graphs in the network and their stability ratio. The latter is measured



**Figure 6.** (a) Whole network: sub-graph abundance (squares) compared with the stability ratio (triangles). Sub-graphs are ordered according decreasing stability ratio. Elements of the Jacobian matrix are drawn from a standard normal distribution. (b) The same test for the GSCC.

over a sampling of 1 million trials per sub-graph. Entries of the Jacobian matrix were drawn from a standard normal distribution: this allows, in contrast, to the approach employed by Prill *et al*, for positive value for the diagonal elements as well. A positive diagonal element in the present case corresponds to autophosphorylation (kinases phosphorylating themselves, which appears to happen quite frequently [28]). A negative diagonal value corresponds to the decay of the phosphorylated state. As it was observed for many other networks [25], the most commonly encountered sub-graphs are those with high stability ratio (see figure 6(a)).

Cyclic sub-graphs are generally less stable, because feedback loops with a potential positive return effect will reduce the stability ratio. Moreover, sub-graphs with a small number of edges are less likely to include cycles. As a consequence, the correlation between sub-graph abundance and stability ratio is not surprising at all, because the most stable sub-graphs are those with less edges and the HSPN is quite sparse. The same analysis applied to the giant strongly connected component (GSCC) provides less intuitive results (see figure 6(b)). The GSCC average degree approximately is similar to that of HPSN, but the cut-off in the sub-graph distribution of GSCC is less dramatic. If stability ratio of sub-graphs affects to a certain extent their distribution, as suggested by Prill *et al* [25], this effect is less evident for the GSCC than for the whole network, though the general trend is the same. Still, it could be argued that the component at the core of the network should not be shaped by evolution for stability, but rather for more ‘exotic’ dynamics through which it transforms signals from IN to OUT components.

## 6. Concluding remarks

In this work, we show that the analysis of directed networks can greatly improve the understanding of the underlying phenomena with respect to simple undirected networks. In fact, the loss of direction information relies on the silent assumption that undirected edges establish communication in both directions. This is, of course, a wrong assumption, since in PSNs a clear direction of signal flow is defined and can strengthen the analysis by encompassing all the available information. Such a distinction is crucial as, for example, ‘hubs’ with mostly outgoing edges will be functionally completely different from hubs with mostly incoming edges, or nodes with a high number of both (see section 4). Some topological properties of HPSN have been characterized before [5] considering an undirected version of the network. For instance, the joint degree distribution has been fit by a power law, while we

show it is a mixture of a truncated-power-law in-degree and a roughly power-law out-degree distribution. Linding *et al* [5] found a decreasing clustering coefficient distribution, which may indicate a hierarchical topology [12]. We showed here that this behaviour still holds for both upstream and downstream neighbourhoods. It should be pointed out that, as a consequence of ignoring the edge direction in a network like this—largely dominated by target nodes (1790 targets versus 69 sources)—the findings are actually explained only by the target features. Concerning the degree and clustering distributions, analysing the undirected version of the HPSN is analogous to analysing the 1790 targets with some added ‘noise’ due to the sources. We pointed out several growing network mechanisms that could give rise to the observed in-degree distribution. As an alternative to such mechanisms, other authors proposed a purely chemical-based model to explain the degree and clustering coefficient distributions in protein interaction networks [29]. They then proposed their model as a suitable null model when looking for evolutionary traces. However, it is not obvious to see how their results extrapolate to the directed protein signalling network we have analysed here. Directed sub-graph analysis and dynamical characterization of the HPSN show interesting features which are worthwhile of further investigation. The distribution of 3-node sub-graph exhibits significant differences with well-known biological networks (e.g. the feed-forward loop is under-represented), which can be partly explained by the topological structure and biological meaning of the HPSN. The stability of the HPSN is driven by the GSCC since nodes belonging to the IN and OUT components contribute only trivially to the spectrum of the Jacobian matrix. The fact that the GSCC has small size, offers the rare opportunity to study the dynamical behaviour of this entire large biological network by the statistical approach outlined in section 5. A detailed investigation on dynamical properties of GSCC is the subject of ongoing research.

The HPSN is a part of a much larger regulatory network, including gene expression—a process that occurs at a slower timescale with respect to protein interaction. The analysis on the HPSN by itself, outside the context of the global human regulatory system, will provide insights into the organization of information processing in cells at least at the fast timescale. For a complete understanding of information processing by human cells the HPSN should be merged with a human transcription regulatory network. However, while parts of the latter network are known [30–32], the genome-wide structure is still in the process of being elucidated. In addition, the HPSN should be extended to include other types of chemical protein modifications, but these are still not known on a proteome-wide scale either.

## Acknowledgments

We thank Ralf Steuer for discussions on the dynamical analysis and Sergio de la Fuente van Bentem for discussions about protein signalling networks. We are grateful to Rune Linding for creating and sharing the HPSN. We kindly acknowledge the support of Regione Autonoma della Sardegna and GM thanks Sardegna Ricerche for fundings.

## References

- [1] Vidal M 2005 Interactome modeling *FEBS Lett.* **579** 1834–8
- [2] Barabasi A L and Oltvai Z N 2004 Network biology: understanding the cell’s functional organization *Nat. Rev. Genet.* **5** 101–13
- [3] Bork P *et al* 2004 Protein interaction networks from yeast to human *Curr. Opin. Struct. Biol.* **14** 292–9
- [4] Pieroni E, de la Fuente van Bentem S, Mancosu G, Capobianco E and de la Fuente A 2008 Protein networking: insights into global functional organization of proteomes *Proteomics* (submitted)
- [5] Linding R *et al* 2007 Systematic discovery of *in vivo* phosphorylation networks *Cell* **129** 1415–26

- [6] Ubersax J A and Ferrell J E Jr 2007 Mechanisms of specificity in protein phosphorylation *Nat. Rev. Mol. Cell Biol.* **8** 530–41
- [7] Guelzim N *et al* 2002 Topological and causal structure of the yeast transcriptional regulatory network *Nat. Genet.* **31** 60–3
- [8] Mancosu G *et al* 2008 Characterization of the global functional organization of a Yeast gene network (in preparation)
- [9] Albert R and Barabasi A L 2002 Statistical mechanics of complex networks *Rev. Mod. Phys.* **74** 47–97
- [10] Amaral L A *et al* 2000 Classes of small-world networks *Proc. Nat. Acad. Sci. USA* **97** 11149–52
- [11] Watts D J and Strogatz S H 1998 Collective dynamics of ‘small-world’ networks *Nature* **393** 440–2
- [12] Ravasz E and Barabasi A L 2003 Hierarchical organization in complex networks *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **67** 026112
- [13] Albert R 2005 Scale-free networks in cell biology *J. Cell Sci.* **118** 4947–57
- [14] Newman M 2003 The structure and function of complex networks *SIAM Rev.* **45** 167–256
- [15] Broder A *et al* 2000 Graph structure in the web *Comput. Netw.* **33** 309–20
- [16] Ma H W and Zeng A P 2003 The connectivity structure, giant strong component and centrality of metabolic networks *Bioinformatics* **19** 1423–30
- [17] Batagelj V and Mrvar A 2003 Pajek—analysis and visualization of large networks *Graph Drawing Software* ed Jünger M and Mutzel P (Berlin: Springer) pp 77–103
- [18] Capocci A *et al* 2006 Preferential attachment in the growth of social networks: the internet encyclopedia Wikipedia *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **74** 036116
- [19] Zhao J *et al* 2006 Hierarchical modularity of nested bow-ties in metabolic networks *BMC Bioinf.* **7** 386
- [20] Dorogovtsev S N, Mendes J F F and Samukhin A N 2001 Giant strongly connected component of directed networks *Phys. Rev. E* **64** 025101
- [21] Maslov S and Sneppen K 2002 Specificity and stability in topology of protein networks *Science* **296** 910–3
- [22] Maslov S, Sneppen K and Zaliznyak A 2004 Detection of topological patterns in complex networks: Correlation profile of the internet *Physica A* **333** 529–40
- [23] Hartwell L H *et al* 1999 From molecular to modular cell biology *Nature* **402** C47–52
- [24] Milo R *et al* 2002 Network motifs: Simple building blocks of complex networks *Science* **298** 824–7
- [25] Prill R J, Iglesias P A and Levchenko A 2005 Dynamic properties of network motifs contribute to biological network organization *PLoS Biol.* **3** e343
- [26] Wernicke S and Rasche F 2006 FANMOD: a tool for fast network motif detection *Bioinformatics* **22** 1152–3
- [27] Steuer R *et al* 2006 Structural kinetic modeling of metabolic networks *Proc. Nat. Acad. Sci. USA* **103** 11868–73
- [28] Ptacek J *et al* 2005 Global analysis of protein phosphorylation in yeast *Nature* **438** 679–84
- [29] Deeds E J, Ashenberg O and Shakhnovich E I 2006 A simple physical model for scaling in protein–protein interaction networks *Proc. Nat. Acad. Sci. USA* **103** 311–6
- [30] Rodriguez-Caso C, Medina M A and Sole R V 2005 Topology, tinkering and evolution of the human transcription factor network *FEBS J.* **272** 6423–34
- [31] Boyer L A *et al* 2005 Core transcriptional regulatory circuitry in human embryonic stem cells *Cell* **122** 947–56
- [32] Odom D T *et al* 2006 Core transcriptional regulatory circuitry in human hepatocytes *Mol. Syst. Biol.* **2** 0017