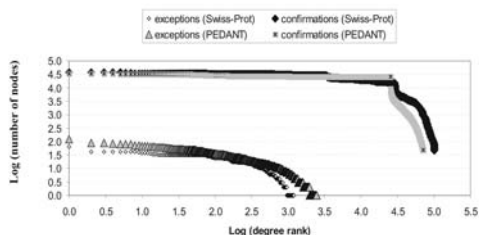


## Annotation rule networks

A convenient way to capture relationships between annotated proteins is via a network in which nodes represent protein entries, and edges join proteins according to certain commonalities in terms of association rules applicable to their annotation. Specifically, in this work we considered two such networks. In the first case two protein nodes were connected by an edge if they both were exceptions to the same strong association rule (with strength from [0.99; 1). In the second case, two proteins needed to satisfy the same perfect (with strength equal 1) association rule to be connected. We refer to the resulting networks as exception network and confirmation network, respectively.

To characterize the networks obtained in this fashion we used two common measures: degree distribution (Barabasi & Oltvai, 2004) and clique size distribution. The former measure defines the fraction of nodes possessing a given number of links while the latter measure indicates how many cliques, or sets of nodes that are all pairwise adjacent, with a given number of nodes are present on the network. We visualized these network properties using a range distribution, *i.e.* the value of the feature with respect to the rank of this value in the list of feature values in the descending order, instead of an ordinary distribution with respect to a feature under study, *i.e.* the relation between the value of the feature and the number of objects with this value. It is easy to show that these two types of plots are equivalent, but the one used is much more illustrative due to its monotony and lower noisiness. We used log-log scale for these plots.

Both networks display a similar behavior and none of them is scale free as demonstrated in Figure I in which we show



**Fig. I.** Rank distribution of the node degree for the networks of exceptions and the networks of confirmations for Swiss-Prot and PEDANT databases

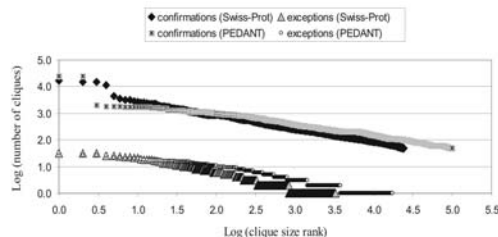
the rank distributions of the node degree in log-log coordinates. These distributions should be linear for scale-free networks; it is clear that this is not the case for both networks presented.

Another natural parameter associated with these networks: the size distribution of the cliques formed by the sets of proteins satisfying or breaking a certain rule. In the exception network the cliques are rather small, with 96.4% of the cliques in Swiss-Prot and 97.7% in PEDANT containing five or less proteins, which is expected as rules with a larger number of exceptions will have lower strength and will not be considered in this network. The clique size in the confirmation networks is simply the coverage of perfect rules (of strength 1). In this work it always exceeds 50 because rules with smaller coverage were not considered at all, and the maximal observed clique size is 16801. Notably, both

clique size distributions follow the power law (Figure II), as is common for protein interaction networks, occurrence of domains in genomes, and many other types of biological data (Barabasi & Oltvai, 2004).

Another interesting parameter associated with the networks is the number of cliques to which a protein belongs. It cannot be described by any simple statistical law (data not shown). The outliers of these distributions are important from the viewpoint of our approach, especially in the case of the exception network, because they represent protein entries that contradict many strong rules simultaneously. The protein associated with the largest number of confirmations is P30530 (precursor of tyrosine-protein kinase receptor UFO from human) that satisfies 1799 perfect rules. Three proteins, P55144, P55146, Q06418 (precursors of tyrosine-protein kinase receptor TYRO3 from mouse, rat and human, respectively), confirm 1773 rules, and the third most prominent protein is P57097 (precursor of proto-oncogene tyrosine-protein kinase MER from rat.) satisfying 1742 rules.

The most “disobedient” protein entry in Swiss-Prot is P00545, tyrosine-protein kinase transforming protein fms. Its annotation contains exceptions to 409 different association rules in the strength range between 0.95 and 1, with 286 of them due to the fact that this protein is the only non-precursor (consequently, without a signal peptide) among many tyrosine kinases with very similar annotation; this case thus does not constitute an error. Further 104 rules contained the feature “ACT\_SITE” in their RHS that was erroneously omitted in annotation of this entry, although it can easily be reconstructed by similarity (data not shown). The second and third top proteins are Q9LYN8, precursor of Leucine-rich repeat receptor protein kinase EXS, contradicting 321 rules, and P42159 (class II receptor tyrosine kinase) which is an exception for 305 rules. The former contradicts 286 rules with “N-LINKED”, “CARBOHYD” or “Glycoprotein” in their RHS, features that can be reconstructed by similarity. Furthermore, 40 other rules with Q9LYN8 having exceptions contain the InterPro domain IPR001245 which was falsely (according to InterProScan, <http://www.ebi.ac.uk/InterProScan/>) assigned to this protein entry up to release 47.0. The latter protein is the only one annotated in mature form in its group. Consequently, 214 rules with the keyword Signal or feature SIGNAL in their RHS and the exception in this protein are correct, whereas the remaining 91 rules for P42159 lead to keyword “Glycoprotein” that have been omitted in its annotation up to the release 46 of Swiss-Prot.



**Fig. II.** Distribution of the clique size for the networks of exceptions and the networks of confirmations for Swiss-Prot and PEDANT databases

## REFERENCE

Barabasi AL, Oltvai ZN. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet.*, 5, 101-13.