Local topological signatures for network-based prediction of biological function

Wynand Winterbach[1,2], Piet Van Mieghem[1], Marcel Reinders[2,3,4],
Huijuan Wang[1], and Dick de Ridder[2,3,4]

[1] Network Architectures and Services Group and
[2] Delft Bioinformatics Lab
Department of Intelligent Systems, Faculty of Electrical Engineering, Mathematics
and Computer Science, Delft University of Technology, P.O. Box 5031, 2600 GA Delft,
[3] Netherlands Bioinformatics Center
6500 HB Nijmegen, The Netherlands,
[4] Kluyver Centre for Genomics of Industrial Fermentation
2600 GA Delft, The Netherlands,
e-mail: w.winterbach@tudelft.nl

Abstract. In biology, similarity in structure or sequence between molecules is often used as evidence of functional similarity. In protein interaction networks, structural similarity of nodes (i.e., proteins) is often captured by comparing node signatures (vectors of topological properties of neighborhoods surrounding the nodes).

In this paper, we ask how well such topological signatures predict protein function, using protein interaction networks of the organism *Saccharomyces cerevisiae*. To this end, we compare two node signatures from the literature – the graphlet degree vector and a signature based on the graph spectrum – and our own simple node signature based on basic topological properties.

We find the connection between topology and protein function to be weak but statistically significant. Surprisingly, our node signature, despite its simplicity, performs on par with the other more sophisticated node signatures. In fact, we show that just two metrics, the link count and transitivity, are enough to classify protein function at a level on par with the other signatures suggesting that detailed topological characteristics are unlikely to aid in protein function prediction based on protein interaction networks.

1 Introduction

To what extent does structure determine function in biology? Evolutionary principles have shown function and structure to be well correlated in genes with common evolutionary ancestors, allowing biologists to infer functions of proteins or genes based on their sequence *homology* (i.e., similarity) with other proteins or genes. With the arrival of network biology [1], homology was extended to take not only sequence similarity into account but also similarity of molecular interactions. These interactions can be either direct (physical) or indirect (functional). In other words, the manner in which a protein (or gene) is connected to other proteins in interaction networks matters. These other connecting proteins can be chosen in many ways, although the most common approach is to consider a network neighborhood centered around a protein in question, including all proteins and links within a fixed number of hops. Structural similarity of network neighborhoods is determined by comparing their *topological* properties. Typically, these properties are represented as a vector, known as a *topological signature*.

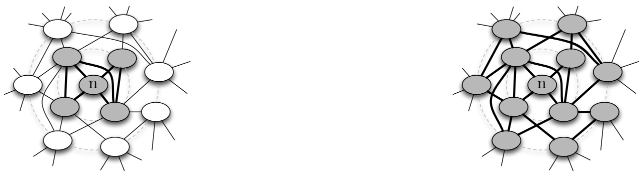
Topological signature similarity has been used as a measure of functional similarity between proteins in several algorithms aimed at the discovery of homology relations between proteins [2–4]. Although topological similarity and amino acid sequence similarity are typically both used to determine homology [2, 4], some of these algorithms perform well using only topological similarity [3, 4]. Researchers have also used topological similarity to predict relations other than homology, in effect assuming that structural similarity implies similarity of biological traits in proteins not necessarily related by evolution. Involvement in cancer (a phenotype) was found to be encoded in topological similarity [5] and even general protein function appears to be encoded in topology [6]. Given this predictive quality, the key question is thus: how exactly does local topology reflect function, and what signatures best capture local topology?

In this paper, we set out to answer these questions in a specific context, i.e. the prediction of protein function by means of node signatures in various protein interaction networks of the organism *Saccharomyces cerevisiae*. Topological signatures in the literature capture a lot of topological detail; in this paper we investigate the extent to which this detail improves protein function prediction (if at all). To this end, we study two such signatures – the graphlet signature of Milenković and Pržulj [6] and a signature based on the normalized Laplacian spectrum of a network [4] – as well as a simple node signature of our design. Predictive power of the signatures is determined by how well they discriminate between proteins with a given biological function and those without the function. To this end we use support vector machines, treating topological signatures as feature vectors and biological labels as classifier labels. Note that our aim is not the construction of an optimal protein function classifier, as for that purpose one would include many other types of data; rather, we use prediction accuracy as a measure to explore the relation between local topology and function.

2 Methods

2.1 Topological signatures

In the remainder of the text, G refers to a network (usually an interaction network), n to an arbitrary node of G and N the number of nodes in G. A k-neighborhood G_n^k of a node n is an induced subnetwork of G on the set of nodes encompassing n and all nodes within k hops of n (a subnetwork is induced when two nodes in the subnetwork are connected by a link if, and only if they are connected in G). The subnetwork G_n^1 spanned by the gray nodes and bold links in Figure 1(a) is a 1-neighborhood of n, whilst the subnetwork G_n^2 spanned by the gray nodes and bold lines in Figure 1(b) is a 2-neighborhood of n.



(a) G_n^1.

(b) G_n^2.

Fig. 1. Two neighborhoods of n: (a) G_n^1 and (b) G_n^2.

Graphlet signature: Graphlets are small, connected, induced subnetworks, as illustrated in Figure 2. The graphlet degree of a node n can be regarded as a generalization of its degree: the number of graphlets of a specific type $(X_1, X_2, \ldots)$ that contains n (the degree is the number of X_1 subgraphs containing n). A graphlet signature (also graphlet degree sequence [6]) generalizes the graphlet degree by including counts for all of the subnetworks in Figure 2.

To simplify exposition, we first construct a graphlet signature containing only the numbers of subnetworks X_1, X_2 and X_3 (Figure 2) that contain n. Such a signature can be represented as a vector of three integers. However, X_2 is not symmetrical, as the white node is structurally different from the two black nodes (which are interchangeable). We distinguish cases in which n takes the role of the white node from cases in which n takes the role of the black nodes. Thus, two counts for X_2 are maintained (one for each kind of node), leading to a signature vector of four integer components: one for X_1, two for X_2 and one for X_3 (vector indices are shown next to one node of each color).

The full graphlet signature is constructed by extending the construction above to the rest of the subnetworks in Figure 2. In total, the signature vector has 73 components (vector indices appear next to nodes). The largest subnetworks in Figure 2 have five nodes and therefore the graphlet signature is computed on 4-neighborhoods. The larger subnetworks in Figure 2 contain induced copies of smaller subnetworks (e.g., X_{30} contains X_9, X_3 and X_1), so that the components of the graphlet signature are not independent. Milenković and Pržulj [6] devised a weighting scheme to reduce this effect. We reweigh graphlets according to their method. Graphlet signatures were computed using code adapted from the original version of GraphCrunch [7].

Spectral signature: We assume that the nodes in G are labeled with numbers 1 through N. The *adjacency matrix* A of G is an $N \times N$ matrix in which $a_{i,j} = 1$

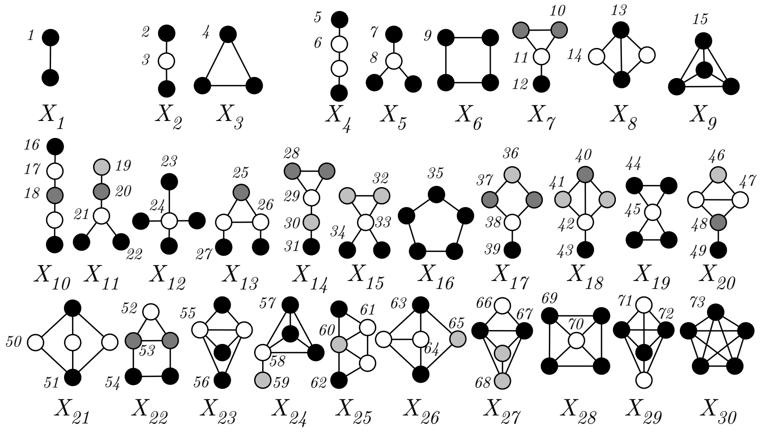


Fig. 2. All non-isomorphic undirected networks (graphlets) with up to five nodes. For a given node n in a network G, Milenković & Pržulj [6] count how many times each of these networks includes n and appears as an induced subnetwork in G in order to construct a graphlet signature for n.

if the nodes i and j are connected by a link and $a_{i,j} = 0$ otherwise. The degree matrix Δ of G is a matrix in which $a_{i,i}$ equals the degree of node i and $a_{i,j} = 0$ if $i \neq j$. The *normalized Laplacian* is defined as $Q_{\text{norm}} = I - \Delta^{-1/2} A \Delta^{-1/2}$. The *spectrum* of Q_{norm} is its set of N eigenvalues. All eigenvalues of Q_{norm} fall within the range of $[0, 2]$.

In general, two different neighborhoods have different numbers of nodes and therefore spectra of different sizes, making spectra unsuitable as feature vectors. We derive feature vectors by computing histograms of the spectra [4]. Histograms with 20 bins are computed on the range $[0, 2]$, showing why the normalized Laplacian spectrum is preferred over the non-normalized version.

Simple metric signature: Our own simple metric signature serves as a baseline. It contains four very simple topological properties of neighborhoods: 1) number of nodes, 2) number of links, 3) link density and 4) transitivity (the ratio of triangles to connected node triplets).

Multi-resolution signatures: One way to compute the spectral and simple metric signatures is to choose a fixed k and to compute the signatures on all k-neighborhoods. By focusing on fixed k, one may miss topologically distinguishing features at other "resolutions", i.e., other values of k. We construct "multi-resolution" versions of the spectral and simple metric signatures respectively by concatenating signatures of G_n^1, G_n^2 and G_n^3 for a given node n; henceforth we shall only consider these "multi-resolution" versions of the signatures. The graphlet signature is already “multi-resolution” in the sense that its component graphlets span G_n^1, G_n^2, G_n^3 and G_n^4.

A combined signature: Finally, we consider a signature that combines the previous signatures by simply concatenating the 1) graphlet signature, 2) the multi-resolution spectral signature and 3) the multi-resolution simple metric signature.

2.2 Datasets

Molecular networks: All of the networks considered in this paper are protein interaction networks for the organism *Saccharomyces cerevisiae.* We have collected seven such networks, derived from four primary sources. Kim & Marcotte [8] provide two protein interaction networks, the first a high-quality literature-curated network and the second a high-throughput network. Yeastnet [9] provides several datasets with yeast protein interactions of which we downloaded the literature-curated dataset (denoted “LC” on the website) and the yeast 2-hybrid high-throughput dataset (“HC”). These two pairs of networks were selected because each pair contains a literature curated network and a high-throughput network, thereby providing insight into the impact of network quality on classification performance.

Our remaining two datasets are due to Krogan [10] and von Mering [11]. Both of these were used by Milenković & Pržulj [6] to test how well their graphlet signature approach fared in predicting protein function. We used the same two subsets of the von Mering dataset: “von Mering” contains the first 11000 protein interactions (of high-, medium- and low-confidence), whilst “von Mering core” contains all high-confidence interactions of the original dataset.

Biological labels: Like Milenković and Pržulj [6], we used the MIPS protein annotations [12] as biological labels. MIPS annotations are hierarchical and have the form “xx.yy.zz. . . ” where the letters denote two-digit biological categories. A protein may be annotated with multiple such annotations. The left-most category (“xx”) gives the general protein function; each following two-digit category is a refinement (“yy” and “zz”). In this paper, we consider only general protein functions, of which there are 27 in the MIPS database.

2.3 Classification

Classification is performed using support vector machines (SVMs). There are numerous biological categories in the MIPS database and a protein may be annotated with any number of these categories. Since SVMs are binary classifiers, we use a one-versus-all strategy whereby we train a classifier for each biological category. Classifier performance is measured using the area under the curve (AUC) of the receiver operator curve (ROC) of a classifier. All classifier-related work was performed using Scikit-learn [13].

The radial basis function (RBF) kernel was used to train all SVMs. To reduce the impact of experimental omissions and noise, we only compute signatures on nodes whose degrees are at least 3 and that have at least one MIPS annotation. Furthermore, to ensure the presence of enough positive instances in both testing and training sets, biological labels that appear in less than 20 nodes are not considered for classification training.

Training regime: For each topological signature type, for each network, for each biological function, a double cross validation training loop is performed [14]. The "outer" loop is a four-fold loop in which the training set contains 75% of the dataset whilst the testing set contains 25% of the dataset. For a given network and biological function, the folds are fixed, meaning that classifiers are trained on the same training samples for all topological signatures. Classifier performance is expressed as a combination of the mean and standard deviation of the four AUC values associated with the four outer folds.

The "inner" loop is responsible for finding the classifier with the best classification performance on the training set received from the "outer" loop. SVM classifiers using the RBF kernel require two parameters: a cost C (for penalizing incorrectly classified instances) and the RBF radius γ. These are optimized by walking along a grid of parameter pairs and training a classifier for each pair. Each grid point (i.e., parameter pair) is evaluated using the average AUC of a five-fold cross-validation loop. The parameters with the best AUC score are thus considered optimal. At the start of the "inner" loop, both the training and testing sets are centered and scaled using the center and variance of the training set. The graphlet signature is reweighed after this point using the weighting scheme of Milenković and Pržulj [6] as mentioned earlier in the paper (if reweighing is applied beforehand, it would be removed by the scaling step).

As grid searches are expensive, we first perform a parameter search on a coarse grid, followed by a second search on a fine grid around the optimal parameters found in the first search. The coarse grid is given by the Cartesian product $\mathcal{C} \times \mathit{\Gamma}$ of costs $\mathcal{C} = \{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{15}\}$ and RBF radii $\mathit{\Gamma} = \{2^{-15}, 2^{-13}, 2^{-11}, \ldots, 2^{3}\}$. The optimal parameter pair (C, γ) discovered on $\mathcal{C} \times \mathit{\Gamma}$ is then used to specify a fine grid $\mathcal{C}' \times \mathit{\Gamma}'$ where $\mathcal{C}' = \left\{2^{\log_2 C - 2 + i/2} \,\middle|\, i \in \{0, 1, \ldots 8\}\right\}$ and $\mathit{\Gamma}' = \left\{2^{\log_2 \gamma - 2 + i/2} \,\middle|\, i \in \{0, 1, \ldots 8\}\right\}$.

3 Results and Discussion

Using the training regime described in the Methods section, we have computed, for each topological signature, for each network, for each biological function, the average classifier performance as well as its standard deviation. As this is a large amount of data, we have condensed the results into Figure 3(a) which shows, for a given topological signature and biological function, classification performance averaged over all networks, except for the high-throughput Yeastnet network. This dataset proved to be too small and gave poor, noisy classification results for all topological signatures. Figure 3(a) contains only those biological functions that appear in all the datasets. We also plotted the classification results for one high-quality dataset, the literature-curated Yeastnet dataset, in Figure 3(b). The trends in Figure 3(a) are broadly similar in all of the networks although classification performance is generally lower than in Figure 3(b).

What stands out most from both Figure 3(a) and Figure 3(b), is that topology is, in general, a weak predictor of biological function. However, the mean AUC values are all above 0.5, showing that topology does encode a certain amount of information about biological function (the statistical significance of the mean AUC values being larger than 0.5 was tested using the *t*-test; in the majority of cases – and in all cases involving the biological categories "metabolism", "transcription", "protein synthesis" and "protein fate" – the associated *p*-values are below 0.05). The overall differences between Figure 3(a) and Figure 3(b) can be explained by differences in network quality and network size: quality affects classifier performance whilst network size affects its variance (network sizes are given in Table 1). The high-throughput networks contain the most noise and are therefore associated with worse classification performance.

At the level of biological categories both Figure 3(a) and Figure 3(b) show big differences in classification performance. The number of positive instances associated with a biological category (see Table 1) is weakly correlated with classifier performance, partly explaining the differences. Biology offers a possible explanation for the high AUC values associated with the labels "Transcription" and "Protein Synthesis": transcription and synthesis are both processes driven by permanent protein complexes rather than temporary groups of proteins (as found in many other processes). Thus, nodes with these functions tend to find themselves in densely connected clusters more often than other nodes.

	Metabolism	Energy	Cell cycle & DNA processing	Transcription	Protein synthesis	Protein fate	Protein w. binding function	Regulation of protein function	Cellular transport	Cellular communication	Cell rescue & defense	Environment interaction	Cell fate	Development	Biogenesis	Cell type differentiation	Number of unique nodes
Kim & Marcotte, HT	271	63	377	481	130	347	399	62	207	46	123	94	80		220	128	1123
Kim & Marcotte, LC	452	84	674	676	149	655	652	157	469	134	235	221	192	35	458	288	1933
Krogan	321	81	423	483	183	378	405	70	205	61	148	115	87		277	134	1281
von Mering core	102	25	75	158	102	88	130		54		22	29	26		84	48	371
von Mering	471	120	231	382	289	295	369	49	193	39	114	99	68		222	104	1307
Yeastnet, HT	96	22	110	90		112	97	26	96	24	52	44	52		99	63	353
Yeastnet, LC	442	82	618	630	207	637	645	142	580	124	222	239	187	39	444	281	2006

Table 1. The number of positive instances for various combinations of network and biological function (i.e., proteins having given biological functions).

Both overall classification performance, as well as performance associated with individual biological categories are dependent on the way in which biological categories are defined. Some categories are more general than others (for example, "Development" includes proteins engaged in diverse functions, whereas "Transcription" is a more specific function), contributing to differences in classification performance between categories. When the categories are too general, overall classification performance suffers as classifier inputs become difficult to distinguish. We have performed experiments (data not shown) in which we used two levels of the MIPS labels (labels of the form "xx.yy" rather than just "xx", i.e., more specific categories). Two-level categories led to better classification performance in some cases (notably those associated with transcription) and worse performance in other cases. The culprit is likely a paucity of positive instances associated with many of the two-level labels.

Another salient aspect of Figure 3(a) and Figure 3(b) is that the three topological signatures perform very similarly. We tested whether the AUC values of the individual signatures (i.e., not the combined signature) for each biological category were different, using a one-way ANOVA (Table 2). We consider p-values of 0.05 and below to be statistically significant and find only 10 dataset/function combinations that pass this threshold.

	Metabolism	Energy	Cell cycle & DNA processing	Transcription	Protein synthesis	Protein fate	Protein w. binding function	Regulation of protein function	Cellular transport	Cellular communication	Cell rescue & defense	Environment interaction	Cell fate	Development	Biogenesis	Cell type differentiation
Kim & Marcotte, HT	.39	.95	**.04**	.17	.39	.77	.69	.14	.10	.15	.19	.23	.35		.85	.56
Kim & Marcotte, LC	.42	.91	.10	.06	**.05**	**.00**	.27	.64	**.01**	.61	.74	**.01**	.31	.76	**.05**	.70
Krogan	.94	.08	.34	.13	.26	.20	.07	.12	.47	.90	.91	.07	.43		.18	.32
von Mering core	.75	.55	.14	.08	.26	.82	.56		.79		.92	.53	.87		.53	.97
von Mering	.19	.32	.49	.12	.59	.24	.26	.14	.24	.06	.50	.43	.60		.17	**.04**
Yeastnet, HT	.44	.22	.12	.36		.68	.19	.07	.18	**.04**	.12	**.00**	.45		.69	.70
Yeastnet, LC	.80	.42	.84	.55	.60	.11	.91	.85	**.04**	.23	.93	.62	.63	**.05**	**.01**	.12

Table 2. p-values of one-way ANOVA tests applied to the AUC values of the three topological signatures (graphlet, spectral and simple) for each network and biological function combination. We consider p-values of 0.05 and below to be significant (shown in bold text).

Although the three topological signatures lead to similar classification results, it may be possible that they nevertheless measure different (discriminative) topological characteristics. If this is true, combining the signatures should lead to improved classification performance. However, Figure 3(a) and Figure 3(b) do not support such a conclusion. Thus, in the context of our datasets and classifier, the topological signatures are not complementary.

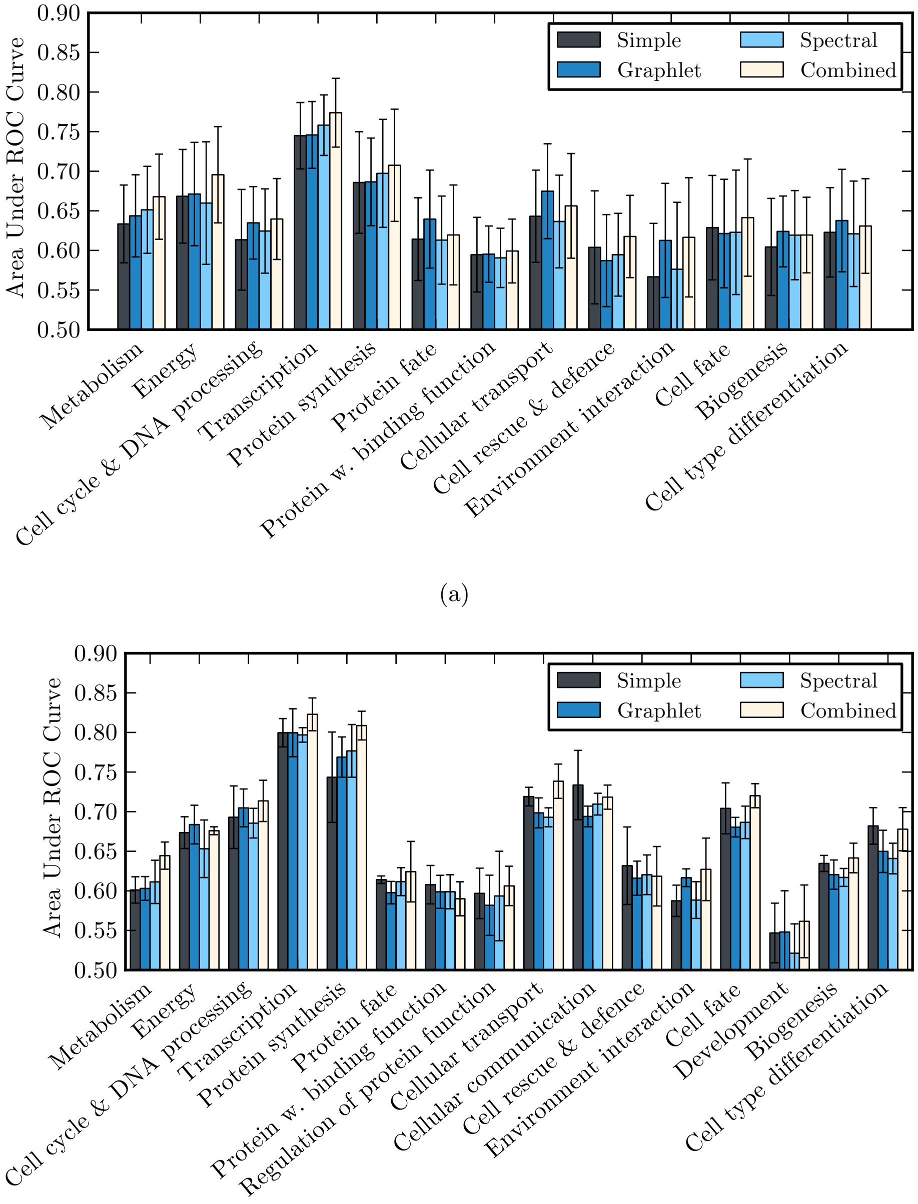
Given that the simple metric signature is competitive with the graphlet and spectral signatures, it is natural to ask whether it cannot be further simplified. We investigated all possible combinations of the four metrics (number of nodes, number of links, density and transitivity) that make up the simple metric signature, constructing 14 simpler signatures: 4 signatures using only one metric each, 6 signatures using pairs of metrics and 4 signatures using triplets of metrics. The mean classification performance of these metrics, taken over all datasets and all biological categories, is shown in Figure 4. The link count L and transitivity T are sufficient for obtaining good classification performance. The implication is that what matters in function prediction in protein interaction networks, is the number of nodes and the “clusteredness” (transitivity). Since proteins of similar function tend to form clusters, their neighborhoods overlap and therefore they share topological characteristics. Apparently, “clusteredness” signatures are unique enough to distinguish similar proteins from other proteins.

4 Conclusion

At the start of this paper, we asked to what extent structure – i.e., topology – determines function in biology. We focused on the use of signatures to express topological properties of neighborhoods surrounding nodes in molecular interaction networks. Our study is motivated by the use of topological signatures as a tool for discovering similar genes or proteins (under the assumption that topological similarity implies functional similarity). We specifically studied the use of such signatures to discriminate between proteins with a given biological function and those without it, using protein interaction networks derived from *Saccharomyces cerevisiae* and support vector machines.

Current node signatures, such as the graphlet signature [6] and signatures based on spectra [4] capture very detailed topological profiles. We compared these with our own topological signature, based on very simple network metrics. For all signatures, classifier performance tended to be weak, implying that topology is, at least for *Saccharomyces cerevisiae* protein interaction networks, a weak predictor of function. However, with the exception of one noisy protein interaction network classifiers performed better than random, showing that topology and function are linked. How much better depends on the functional category considered, with performance particularly strong for transcription and protein synthesis.

Our simple metric signature performed on par with the graphlet and spectral signatures. We also established that the signatures are not complementary for protein function prediction, as a combined signature incorporating all three signatures does not yield better accuracy. Since our simple metric signature captures less topological information than the other signatures, we conclude that fine topological detail is not very useful in the prediction of protein function.



(a)

(b)

Fig. 3. Classification performance of the three topological signatures, as well as a signature that combines the three signatures. (a) Performance of our SVM classifiers averaged MIPS categories present in all datasets (excluding the high-throughput Yeastnet dataset; see text for explanation). Error bars show the standard deviation. (b) Classification performance of the three topological signatures on the literature-curated Yeastnet network [9].

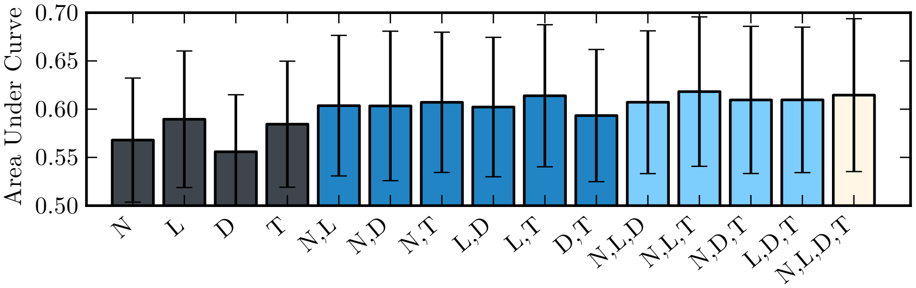


Fig. 4. Classification performance of various combinations of the features used in the simple metric signature averaged over all datasets and all functions. Here, N is the number of nodes (in a neighborhood), L is the number of links, D is the density and T is the transitivity.

Strikingly, performance when using only the link count and transitivity, measures of "clusteredness", is as good as when using the more complex signatures. This is not simply a side-effect of dataset noise, as our simple metric signature performs equally well in the high quality networks.

Our work opens a number of paths for future research. For our conclusions to hold generally, the techniques used in this paper should be applied to other types of interaction networks (for example, co-expression networks and synthetic sick-or-lethal networks) and to networks derived from other organisms. It would be particularly interesting if link count and transitivity are found to be equally determinative in other interaction network types. Finally, it is not yet known how different "resolutions" contribute to signature performance and whether a particular resolution (i.e., k-neighborhoods of a particular k) dominates classification performance.

References

1. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.L.: Hierarchical organization of modularity in metabolic networks. Science **297**(5586) (August 2002) 1551–1555
2. Liao, C.S., Lu, K., Baym, M., Singh, R., Berger, B.: IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics **25**(12) (June 2009) i253–i258
3. Milenković, T., Ng, W.L.L., Hayes, W., Pržulj, N.: Optimal network alignment with graphlet degree vectors. Cancer Informatics **9** (2010) 121–137
4. Patro, R., Kingsford, C.: Global network alignment using multiscale spectral signatures. Bioinformatics (2012)
5. Milenković, T., Memišević, V., Ganesan, A.K., Pržulj, N.: Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. Journal of The Royal Society Interface **7**(44) (2010) 423–437

6. Milenković, T., Pržulj, N.: Uncovering biological network function via graphlet degree signatures. Cancer informatics **6** (2008) 257–273
7. Milenkovic, T., Lai, J., Pržulj, N.: GraphCrunch: A tool for large network analyses. BMC Bioinformatics **9**(1) (2008) 70
8. Kim, W.K., Marcotte, E.M.: Age-Dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. PLoS Computatinal Biology **4**(11) (November 2008)
9. McGary, K., Lee, I., Marcotte, E.: Broad network-based predictability of saccharomyces cerevisiae gene loss-of-function phenotypes. Genome Biology **8**(12) (2007) R258
10. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A., Greenblatt, J.F.: Global landscape of protein complexes in the yeast saccharomyces cerevisiae. Nature **440**(7084) (March 2006) 637–643
11. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417**(6887) (May 2002) 399–403
12. Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M., Mewes, H.W.: The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research **32**(18) (2004) 5539–5545
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12** (2011) 2825–2830
14. Wessels, L.F.A., Reinders, M.J.T., Hart, A.A.M., Veenman, C.J., Dai, H., He, Y.D., van 't Veer, L.J.: A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics **21**(19) (2005) 3755–3762