Protein Cluster Analysis via Directed Diffusion

Yosi Keller[1] and Stephane Lafon[2]

[1]Department of Mathematics, Yale University and [2]Google Inc.

ABSTRACT

Motivation: Graph-theoretical approaches are useful for elucidating the modular compositions of protein-protein interaction networks, which are known to consist of regions of increased network connectivity (clusters), corresponding to known molecular complexes or functional pathways. In this work, we introduce the concept of semi-supervised *directed diffusion* as a graph-based methodology for cluster analysis.

Results: We show that our scheme allows both similarity propagation and cluster boundary detection. It is experimentally verified by analyzing known biological pathways, and we show that it can accurately identify an entire pathway, given only 10%-20% of its proteins. Thus, we submit that directed diffusion is a promising approach for evidence propagation in biological networks and clustering of functional groups.

Availability: A Matlab implementation of the proposed scheme is available at http://pantheon.yale.edu/∼yk253/software/.

Contact: yosi.keller@yale.edu.

1 INTRODUCTION

There is much evidence that topological features of molecular interaction networks are useful for delineating the location and scope of molecular pathways and complexes. Moreover, it has been shown that network density is a useful feature for characterizing network regions consisting of elements, such as proteins, that preferably interact with one another and participate in a common biological function. Previous work dealing with the identification of such regions (commonly called network clusters) often represented molecular interaction networks as graphs. These capture the network elements as nodes and the interactions among the elements as network edges. Such methods can be divided into two categories: unsupervised data-driven schemes in contrast to semi-supervised approaches that incorporate prior knowledge. (Rives and Galitski, 2003) and (Wilkinson and Huberman, 2004) present data driven approaches that perform hierarchical clustering of networks, utilizing different affinity measures and edge weights to identify nodes and edges at the boundary of network clusters. Wilkinson et. al. (Wilkinson and Huberman, 2004) followed the work by (Girvan and Newman, 2002) (GN algorithm) to subdivide disease-specific networks into meaningful gene communities. The GN algorithm starts with a fully connected network, and successively removes edges between compact network clusters. Another approach is discussed by (Pereira-Leal et al., 2004), which uses the TribeMCL method (Enright et al., 2002) to find functional modules in protein interaction networks. They applied the concept of a 'stochastic flow' that alternates between expansion and inflation phases, causing the flow to dissipate within clusters, while eliminating it between clusters. There is a body of research that deals with the delimitation of network clusters from a set of seed nodes that corresponds to a-priory knowledge about a the clusters of interest. Bader and Hogue (Bader and Hogue, 2003) developed MCODE that allows to explore network clusters from specified seed nodes (such as proteins). Bader developed SEEDY (Bader, 2003), an algorithm for finding protein clusters through breath-first outward traversal from some known seed genes.

Algorithms based on random walks and physical computational models are of particular interest in the context of this work. An approach based on the thermodynamical properties of an inhomogeneous ferromagnet was proposed by Domany and coauthors (M. Blatt and Domany, 1996; Blatt et al., 1997). This approach models the interaction between data points by assigning a spin to each data. Each spin can be in several states. Spins belonging to connected nodes interact and have the lowest energy when they are in the same state. The system (known as the Potts model) is subject to equilibration at nonzero temperature, making spins fluctuate. The concept behind this method is that spins belonging to a highly connected cluster fluctuate in a correlated fashion. By detecting correlated spins, the algorithm can identify nodes belonging to a highly connected area of the graph. This approach was successfully applied to a variety of clustering problems (Getz et al., 2000, 2002; Reichmann et al., 2005; Spirin and Mirny, 2003), and extended in (Getz et al., 2005) to deal with semi-supervised classification, where the classes of certain samples are know a priori. The Potts model is augmented by introducing an external field into the energy of the granular ferromagnet.

A graph theoretic approach to unsupervised data analysis was suggested by Harel and Koren in (Harel and Koren, 2001b,a). They compute the affinity matrix related to the dataset, and row normalize it to form a Markov matrix. A random walk is then initiated at all of the datapoints, and the probabilities induced by it are used as a set of features defined on the dataset. A new graph is then defined based on the inner similarities of this new set of features. This new graph is then clustered by a *Clustering by Separation* approach, where a greedy algorithm iteratively removes edges until the clusters are formed.

The work in (Weston et al., 2004) is of particular interest to us, as it presents a protein ranking algorithm that utilizes an evidence propagation approach similar to ours. The work exploits the entire network structure of affinities among proteins in a sequence database, by performing a diffusion operation on a precomputed, weighted network.

In this work, we follow a related approach for detecting network clusters by taking a closer look at spectral graph theory. This approach was shown to provide an efficient and fundamentally sound framework for data analysis and statistical clustering (Shi and Malik, 2000). In such schemes multidimensional data is analyzed by forming a graph of interactions/distances between the data elements, such as nodes in a molecular interaction networks. A data set can then be clustered by studying the spectral properties of the graph

Laplacian and computing the so-called diffusion distances over the data manifold (Coifman et al., 2005b).

We present a semi-supervised approach for delineating the scope of protein clusters given a few samples nodes (seed proteins) within it. We call this approach *directed diffusion*, as we direct our analysis towards a particular cluster instead of analyzing the entire data set. Our approach is based on propagating the diffusion from the seeds, using a Markov matrix and identifying the target cluster's boundaries by the discontinuities in the diffusion values. Let M be a Markov matrix, then $p_{n+1}^T = (p_n)^T \cdot M$ defines a Markov random walk, while in this work we propagate the diffusion values $\phi_{n+1} = M \cdot \phi_n$ and ϕ_n is not a probability.

The proposed scheme differs from the prior results cited above in several aspects. Compared to unsupervised approaches (M. Blatt and Domany, 1996; Blatt et al., 1997; Bader, 2003), our approach utilizes prior knowledge and does not require the analysis of the entire dataset, which in the case of interaction networks, might prove computationally prohibitive due to their size. Compared to schemes that use Markov random walks, such as the work by Harel and Koren in (Harel and Koren, 2001b,a), our approach propagates diffusions that are shown to have particular analytical properties useful for clustering. In contrast, in (Harel and Koren, 2001b,a), the random walk is used to derive a set of coordinates and the clustering is conducted via a general purpose graph theoretic scheme. In (Weston et al., 2004) the authors propagate diffusions similar to us, but their scheme provides only ranking and not clustering. Last, using a random sampling procedure, our approach naturally handles noisy seeds. This issue was not addressed in prior works and is vital for processing experimental data that is prone to errors.

2 DIFFUSION BASED DATA ANALYSIS

Graph based diffusion (Coifman et al., 2005b) is an approach to analyzing the geometrical properties of data networks represented as graphs. It is particularly appropriate for data sets with a pair-wise similarity function between the data points. Its focal point is to agglomerate the local interactions of the different entities (data points) in order to obtain a global or semi-global embedding.

Suppose we have a data set $E = \{x_1, ..., x_L\}$ and a non-negative and symmetric similarity measure $w(x_i, x_j)$. This results in a symmetric graph, whose L nodes are the data points E with the symmetric affinity matrix $W = \{w(x_i, x_j)\}$. For $x_i \in E$, we form the $L \times L$ matrix M with entries

$$m(x_i, x_j) = \frac{w(x_i, x_j)}{\sum_{j=1}^{L} w(x_i, x_j)}.$$

As $m(x_i) \geq 0$ and $\sum_{j=1}^{L} m(x_i, x_j) = 1$, the matrix M can be viewed as the transition matrix of a random walk (Markov chain) on the data set E. The diffusion framework is based on the idea that *the behavior of this random walk provides some insight on the structure of the data set*. Although the construction of this Markov chain might seem artificial, it can reveal useful information about the organization of the data. Consider a random walker traveling on the graph of the data, jumping from node to node at random. Each time he intends to move, he chooses a destination at random according to $m(x_i, x_j)$. Therefore, from x_i, he is more likely to jump to x_j if the affinity between these nodes is high, and more generally, he will follow paths along which nodes are highly similar (in the sense defined by w). The walker will tend to be trapped in a certain number of regions or wells for a long time, with some rare jumps between these wells. From this observation, it is natural to define a *cluster* in the data set E to be a region where the random walker tends to spend a lot of time. In contrast, *regions of transitions*, or *bottlenecks*, constitute the boundaries of these clusters.

In unsupervised schemes, the random walk is initiated in all of the samples at once, and we can compare two points x and z in E, by comparing the behaviors of two random walks starting at x and z respectively. Let $p_n(x, y)$ denote the probability of being in y at time t when starting at x, then the comparison of the distributions $p_n(x, \cdot)$ and $p_n(z, \cdot)$ can be used for measuring a notion of distance between x and z denoted as the *diffusion distance* (Coifman et al., 2005b,a)

$$D_t^2(x, z) = \sum_{y \in E} \frac{1}{d(y)} \left(p_t(x, y) - p_t(z, y)\right)^2 . \tag{1}$$

where $d(y)$ is the degree of the node y.

In contrast, in this work, we compare *diffusions* rather than probabilities. The points x and z in E are compared by considering the diffusions $\phi_n(y, x)$ and $\phi_n(y, z)$, where $y \in S$, $S \subset E$, is the point where the diffusion is initiated. In a biological context, E can be the set of proteins represented by an interaction network and S might be the cellcycle pathway. This allows us to focus our analysis on S and avoid the computation of the eigenfunctions of the Markov matrix M in Eq. 1. These eigenfunctions are global and embed the entire graph. Therefore, if one only wants to study the neighborhoods of the set S, a large number of these eigenfunctions are needed to achieve the desired resolution. In addition, the computational complexity of computing them might prove prohibitive for large networks.

In order to identify the set S we utilize the analysis given in (Keller, Keller), where it was shown that $\phi_n(y, \cdot)$ is discontinuous across the boundary of S, and this discontinuity is followed by a saddle-like structure corresponding to the cluster S. In the next section we describe how to utilize these notions to derive an unsupervised clustering scheme applicable to proteins networks analysis,

3 EVIDENCE PROPAGATION VIA DIRECTED DIFFUSION

Directed diffusion is a mean to identify the members of S, when only a subset of them $S_o \subset S$ is known. We call these proteins *seeds*. For instance, if the dataset is a collection of E proteins, then S can be a subset of proteins that participate in a common biological pathway (e.g. cell cycle). For that we start by propagating the diffusions initiated from set S_o using Algorithm 1.

Algorithm 1 Computing the probability ϕ_n

```
Let $\phi_0$ be the indicator vector that is zero outside $S_o$, and 1 on
  $S_o$. $n_{\max}$ is a predefined number of time steps.
for $n = 1$ to $n_{max}$ do
    Compute $\widetilde{\phi}_n = (\phi_{n-1})^T M$
    Define $\phi_n$ by $\phi_n(x) = \widetilde{p}_n(x)$ if $x \notin S_0$ and $\phi_n(x) = 1$ if
      $x \in S_0$.
end for
Return $\phi_n$ where $n = n_{\max}$.
```

Applying Algorithm 1 results in a vector of diffusions ϕ_n, in which high values correspond to nodes with high affinity to the set S. In order to cluster the elements of S, we utilize the property of ϕ_n discussed in (Keller, Keller), where it was shown that ϕ_n will be discontinuous across the boundary of the set S, denoted ∂S. Thus, nodes $\{x_i\} \in E$ that are outside of S will have smaller diffusion values ϕ_n, and we denote the value of ϕ_n at the discontinuity point T^*. The set S can be then clustered by

$$\begin{array}{ll} x \in S & \text{if } \phi_n(x) > T^* \\ x \notin S & otherwise. \end{array}$$

Let $N(\phi_n > T) = |\phi_n > T|$ be the number of samples with diffusion values larger than T, then in order to identify the discontinuity T^*, we utilize the observation that $N(\phi_n > T)$ is a monotonically decreasing function of T and is discontinuous (with respect to T) at T^*. Thus, we aim to find the largest discontinuity in $N(\phi_n > T)$ and the saddle point following it. Note that $N(\phi_n > T)$ is a robust measure as it is computed over the entire set E. This is summarized in Algorithm 2.

Algorithm 2 Computing the threshold T^*

```
Apply Algorithm 1 and compute $\phi_n$.
Compute $N(\phi_n > T) = |\phi_n > T|$ for $0 \leq T \leq 1$.
Identify the largest discontinuity
    $T_{\max} = \arg\max \left| \frac{\partial N(\phi_n > T)}{\partial T} \right|$.
Look for $T^*$, the first saddle point of $\left| \frac{\partial N(\phi_n > T)}{\partial T} \right|$ such that
    $T^* > T_{\max}$.
```

3.1 Random sampling in diffusion space

In practice, S_o might contain erroneous seeds S_f such that $S_o = S_f \cup S_t$ but $S_f \notin S$ and $S_t \in S$. A straightforward implementation of Algorithm 1 can not distinguish between true and false seeds, as the diffusion values of the seeds are a-priory set to 1.0. Thus, In order to analyze the seeds themselves, we introduce a random sampling scheme to account for the existence of erroneous seeds. We randomly choose a subset of S_o, and propagate from it using Algorithm 1. This is repeated K times and we denote the randomly chosen subset S_0^k, and the corresponding diffusion vector ϕ_n^k. After K such iterations, we average $\{\phi_n^k\}$ over k and denote the result $\overline{\phi_n}(x)$. For each iteration k, we do not average over the samples S_0^k that were used as seeds. This corresponds to only considering *secondary* evidence for the seeds. Assuming $|S_t| \gg |S_f|$, the affinity is likely to spread faster within S_t than within S_f, and $\overline{\phi_n}(S_t) \gg \overline{\phi_n}(S_f)$ .This approach is described in Algorithm 3.

Algorithm 3 Random sampling

```
1: Given the set of seeds $S_o$, the random sampling ratio $r$ and the
   number of random iterations $K$.
2: for $k = 1$ to $K$ do
3:     Choose a random set $S_0^k \subset S_o$ made of $r \cdot |S_o|$ elements and
       define the function $Q^k(x)$ such that
           $Q^k(x) = \begin{cases} 1 & x \in S_0^k \\ 0 & otherwise \end{cases}$
4:     Apply Algorithm 1 using $S_0^k$ as seeds and denote the result
       $\phi_n^k$.
5: end for
6: Compute the weighted average
       $\overline{\phi_n}(x) = \frac{\sum_{k=1..K} \phi_n^k(x) Q^k(x)}{\sum_{k=1..K} Q^k(x)}.$   (2)
```

4 RESULTS AND DISCUSSION

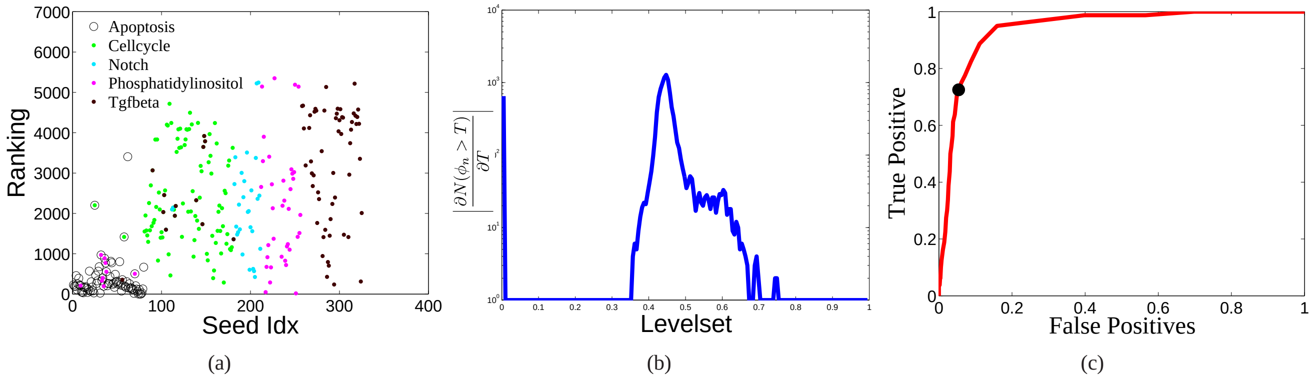
The effectiveness of the directed diffusion was experimentally verified by analyzing the cluster boundaries of the cell cycle (100 proteins), apoptosis (80 proteins) and TGF-beta (69 proteins) pathways. These are known biological pathways that were used as ground-truth.

Our analysis is based on a protein-protein interaction network constructed using the Human Reference Protein Database (HRPD) (Peri et al., 2003). This network provides hand-curated interaction data for several thousand human proteins, and is based on data from the scientific literature. Where possible, it provides external database references for the listed proteins; we discarded HRPD entries without such references, as we were unable to uniquely identify the corresponding proteins. The resulting human protein interaction network consisted of 5537 nodes and is a binary network where nodes are either connected or not. Within this network, we identified protein nodes that belong to the cell cycle, apoptosis and three different signaling pathways (Notch, TGF-beta and phosphatidylinositol signaling pathways). For that, we used the KEGG curated pathway repository (Kanehisa et al., 2004) to obtain the corresponding gene identifiers.

In terms of Section 3, E is the entire set of 5537 proteins encoded by the network. In each experiment we analyzed a single pathway, that acted as the set S. In order to derive the set of seeds S_o, we randomly chose 20% of the samples in S and applied the iterative random sampling described in Algorithm 3 using the random sampling ratio $r = 50\%$. We used 150 random walk steps in Algorithm 1, and the entire experiment was repeated 200 times (i.e. we chose 200 different sets S_o).

For each protein we computed the diffusions according to Step#6 in Algorithm 3 and ranked the proteins according to the descending order of their average diffusions $\overline{\phi_n}(x)$. Thus, proteins with high $\overline{\phi_n}(x)$ values are ranked low. This allows us to verify the validity of Algorithm 1 as a similarity propagation measure, as low ranking are expected to correspond to samples related to the seeds and pathway of interest. For a successful diffusion propagation, we expect the proteins in S to have low rankings. Indeed, the rankings of the apoptosis, cell cycle and TGF-beta pathways (shown in Figs. 1a-3a, respectively), are all considerably lower than the rankings of the other pathways. In contrast, we applied the same procedure to a random set of seeds, and the resulting rankings shown in Fig. 4a, are uniformly spread over the range of 1500-3500.

This exemplifies that proteins clusters with low ranking are not formed by random seeds, and it is evident that the rankings (and



(a)

(b)

(c)

Fig. 1. Analysis results for the Apoptosis pathway that contains 80 proteins. The diffusions were initiated using 20% (16 samples) of the proteins. (a) The ranking of the probabilities ϕ_n of five protein pathways high-to-low. The Apoptosis proteins (black circles) show the lowest rankings and highest probabilities. (b) We detect the largest discontinuity in ϕ_n and the saddle point is detected at $T^* = 0.5$ and overlayed on the ROC curve in (c). This choice of T^* corresponds to an accurate classification of the Apoptosis pathway. In addition, the ROC curve has an area of 0.942.

diffusion values) of the random seeds are indistinguishable from the rest of the network.

Next, we tested the clustering component of our scheme that is described in Algorithm 2. For each pathway we computed the corresponding function $\left|\frac{\partial N(\phi_n>T)}{\partial T}\right|$ and depict it in Figs. 1b-3b. For each of them, we look for the saddle point T^*, which follows the largest maxima of $\left|\frac{\partial N(\phi_n>T)}{\partial T}\right|$. Note that the left-most discontinuity at $T^* = 0$ corresponds to disconnected nodes/proteins.

In order to asses the classification accuracy we computed the ROC curve (Fawcett, Fawcett) for each pathway, shown in Figs. 1c-3c, by choosing a set of thresholds $\{T_i\}$, uniformly spread over the interval $[0..1]$. For each threshold we identify the nodes for which $\phi_n > T_i$ as belonging to the pathway of interest, using the KEGG repository as ground truth. Thus, the ROC curve depicts the classifier's performance with respect to all possible values of T^*, and we mark the threshold T^*, computed according to Algorithm 2, on the ROC curve. An optimal classifier (and value of T^*) will be located as close as possible to the upper-left corner of the ROC curve, where it maximizes the ratio of true positives to false positives. It is evident in Figs. 1c-3c that the chosen thresholds T^* are indeed close to the optimal position. In contrast, in Figs. 4b-4c, where seeds were picked at random, the ROC curve correspond to that of a random classifier (ROC area equal to 0.5) and the saddle point is meaningless.

In order to study the directed diffusions' dependency on the number of seeds, we computed the ROC curves for varying numbers of seeds. It is well known (Fawcett, Fawcett) that a random classifier would yield an area of 0.5, while an optimal classifier yields an area of 1.0. These results are shown in Figs. 5a-5c and it is evident that 10% of the proteins within each pathway suffice to recover the remaining proteins by achieving a ROC area of 0.9. For comparison, we computed the ROC area for a directed diffusion classifier based on random seeds. As expected the average ROC area is 0.5 regardless of the number of seeds, and the real and random classifiers are more than one standard deviation apart.

5 CONCLUSIONS

This work discusses the application of directed diffusion to the analysis of protein interaction data. We presented a task driven approach to protein clustering within large interactions networks, by propagating the diffusion locally using a given set of seed proteins. This enables the detection of protein cluster boundaries, and (using a random sampling approach) overcomes the presence of corrupt seeds. We demonstrated the applicability of our approach by applying it to a protein interaction network. To conclude, we believe that we provided compelling evidence that directed diffusion is an efficient evidence propagation and

6 ACKNOWLEDGMENTS

The author would like to thank Amit Singer (Applied Math department, Yale University) for many helpful discussions during the course of this work, regrading the Fokker-Planck operator. We also thank Michael Krauthammer for the biological datasets and many helpful insights.

REFERENCES

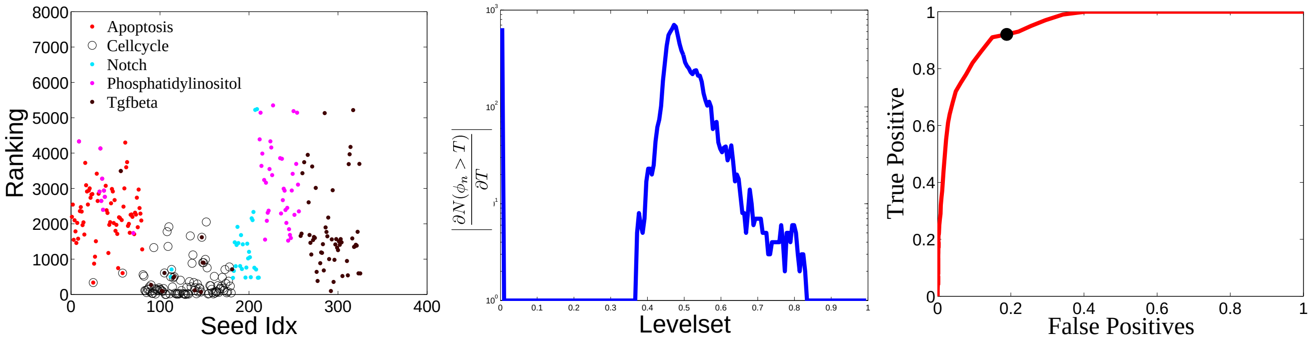
Bader, G. D. and C. W. Hogue (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics 4*(1), 2.

Bader, J. S. (2003). Greedily building protein networks with confidence. *Bioinformatics 19*(15), 1869–74.

Blatt, M., S. Wiseman, and E. Domany (1997). Data clustering using a model granular magnet. *Neural Computation 9*(9), 1805–1842.

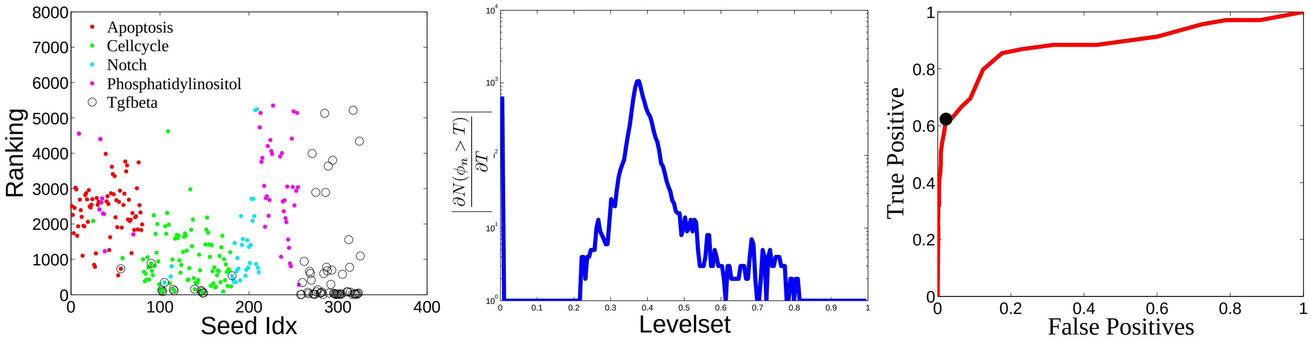
Coifman, R., S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker (2005a). Geometric diffusions as a tool for harmonics analysis and structure definition of data. part 2: Multiscale methods. *Proceedings of the National Academy of Science*. To appear.

Coifman, R. R., S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker (2005b, May). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *PNAS 102*(21), 7432–7437.



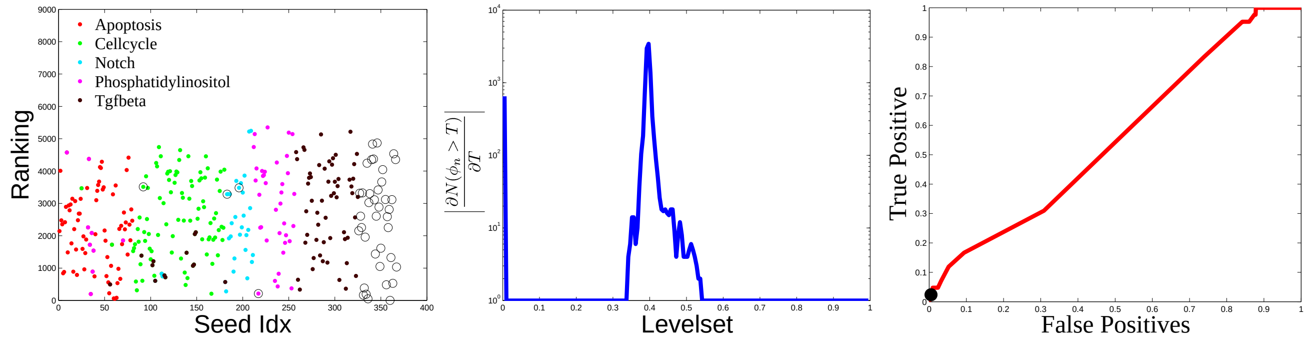
(a) (b) (c)

Fig. 2. Analysis results for the Cell Cycle pathway that contains 100 proteins. The diffusions were initiated using 20% (20 samples) of the proteins. (a) Ranking of the probabilities ϕ_n of five protein pathways high-to-low. The Apoptosis proteins (black circles) show the lowest rankings and highest probabilities. (b) We detect the largest discontinuity in ϕ_n and the saddle point is detected at $T^* = 0.525$ and overlayed on the ROC curve in (c). This choice of T^* corresponds to an accurate classification of the Apoptosis pathway. In addition, the ROC curve has an area of 0.937.



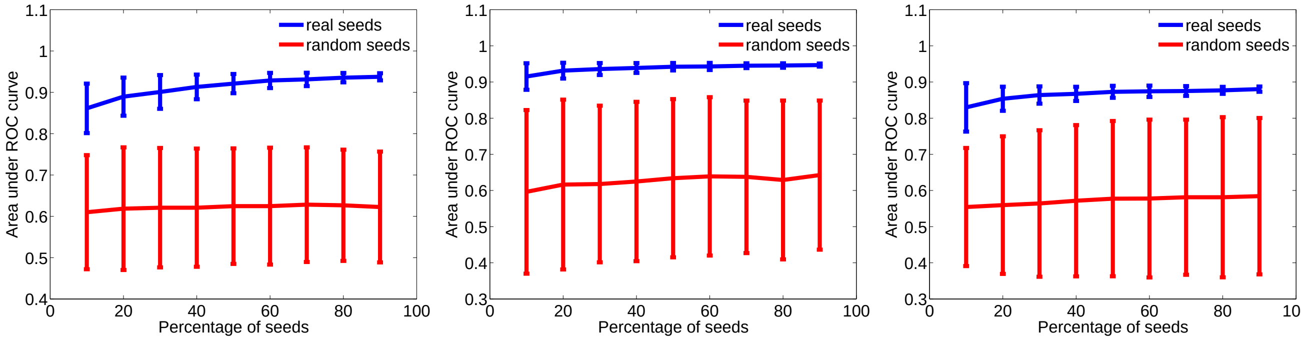
(a) (b) (c)

Fig. 3. Analysis results for the TGF-beta pathway that contains 69 proteins. The diffusions were initiated using 20% (13 samples) of the proteins. (a) Ranking of the probabilities ϕ_n of five protein pathways high-to-low. The Apoptosis proteins (black circles) show the lowest rankings and highest probabilities. (b) We detect the largest discontinuity in ϕ_n and the saddle point is detected at $T^* = 0.46$ and overlayed on the ROC curve in (c). This choice of T^* corresponds to an accurate classification of the Apoptosis pathway. In addition, the ROC curve has an area of 0.883.



(a) (b) (c)

Fig. 4. Analysis results for the a randomly chosen set of seeds containing 84 proteins. The diffusions were initiated using 20% (17 samples) of the proteins. (a) Ranking of the probabilities ϕ_n of five protein pathways high-to-low. The random seeds are spread randomly and not characterized by low rankings as the actual pathways. (b) The saddle point following largest discontinuity is overlayed in (c), but does not correspond to an accurate classification. (c) As expected, the ROC curve corresponds to the output of a random classifier and has an area of 0.55



(a)

(b)

(c)

Fig. 5. ROC results for (a) apoptosis, (b) cell cycle and (c) TGF-beta. In all three cases, the average ROC curve of the directed diffusion classifier, is well above the average ROC curve computed using random seeds. The average ROC area achieved by the directed diffusion is in the order of 90%.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res 30*(7), 1575–84.

Fawcett, T. Roc graphs: Notes and practical considerations for data mining researchers. *Technical Report HPL-2003-4, HP Labs, 2003.*.

Getz, G., E. Levine, and E. Domany (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Science 97*(22), 12079–12084.

Getz, G., N. Shental, and E. Domany (2005, August). Semi-supervised learning - a statistical physics approach. Bonn, Germany.

Getz, G., M. Vendruscolo, D. Sachs, and E. Domany (2002). Automated assignment of scop and cath protein structure classifications from fssp scores. *Proteins* (46), 405415.

Girvan, M. and M. E. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Science 99*(12), 7821–6.

Harel, D. and Y. Koren (2001a). Clustering spatial data using random walks. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 281–286. ACM Press.

Harel, D. and Y. Koren (2001b). On clustering using random walks. In *FST TCS '01: Proceedings of the 21st Conference on Foundations of Software Technology and Theoretical Computer Science*, London, UK, pp. 18–41. Springer-Verlag.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori (2004). The kegg resource for deciphering the genome. *Nucleic Acids Res 32*(Database issue), D277–80.

Keller, Y. Local spectral analysis and its applications. *Submitted for publication*.

M. Blatt, S. W. and E. Domany (1996, April). Superparamagnetic clustering of data. *Physical Review Letters 76*, 32513254.

Pereira-Leal, J. B., A. J. Enright, and C. A. Ouzounis (2004). Detection of functional modules from protein interaction networks. *Proteins 54*(1), 49–57.

Peri, S., R. Navarro, J.and Amanchy, T. Kristiansen, C. Jonnalagadda, and V. Surendranath (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res 13*(10), 2363–71.

Reichmann, D., O. Rahat, S. Albeck, R. Meged, O. Dym, and G. Schreiber. (2005, January). The modular architecture of protein-protein binding interfaces. *Proceedings of the National Academy of Science 102*(1), 57–62.

Rives, A. and T. Galitski (2003). Modular organization of cellular networks. *Proceedings of the National Academy of Science 100*(3), 1128–33.

Shi, J. and J. Malik (2000, August). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 888–905.

Spirin, V. and L. A. Mirny (2003). Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Science 100*(21), 12123–12128.

Weston, J., A. Elisseeff, D. Zhou, C. Leslie, and W. S. Noble (2004). Protein ranking: from local to global structure in the protein similarity network. *Proceedings of the National Academy of Science 101*(17), 6559–6563.

Wilkinson, D. M. and B. A. Huberman (2004). A method for finding communities of related genes. *Proceedings of the National Academy of Science 101 Suppl 1*, 5241–8.