Mining Disease Fingerprints From Within Genetic Pathways

Ahmed Ragab Nabhan, MSc[1,3,4] and Indra Neil Sarkar, PhD, MLIS[1,2,3]
[1]Center for Clinical & Translational Science, [2]Department of Microbiology & Molecular Genetics, and
[3]Department of Computer Science, University of Vermont, Burlington, VT, USA;
[4]Faculty of Computers & Information, Fayoum University, Egypt

Abstract
Mining biological networks can be an effective means to uncover system level knowledge out of micro level associations, such as encapsulated in genetic pathways. Analysis of human disease genetic pathways can lead to the identification of major mechanisms that may underlie disorders at an abstract functional level. The focus of this study was to develop an approach for structural pattern analysis and classification of genetic pathways of diseases. A probabilistic model was developed to capture characteristic components ('fingerprints') of functionally annotated pathways. A probability estimation procedure of this model searched for fingerprints in each disease pathway while improving probability estimates of model parameters. The approach was evaluated on data from the Kyoto Encyclopedia of Genes and Genomes (consisting of 56 pathways across seven disease categories). Based on the achieved average classification accuracy of up to ~77%, the findings suggest that these fingerprints may be used for classification and discovery of genetic pathways.

Introduction

Biological cells have sophisticated information processing systems with highly modular architectures. The flow of information in and between cells can be achieved through a series of biochemical interactions that are composed of a network with a fixed or changing topology. Gaining insight into the operations of cells requires the analysis of components (e.g., genetic material, chemical molecules, and compounds), identifying links (wiring) that represent relations or interactions between components, and discovering information pathways in these networks. Analysis of the structure and dynamics of biological networks plays an important role in understanding architecture and function of biological systems. To level the landscape for a system-based understanding of cellular processes, there has been much previous work in the construction of biological network models, accompanying databases, and development of identification (prediction) algorithms of genetic pathways[1-5].

Network medicine[6] represents one application area where the analysis of biological networks has a potentially direct impact on human health. In this regard, the analysis of genetic pathways may advance knowledge towards an understanding of the molecular underpinnings of the disease process[7-12]. Important questions about complex diseases, such as Alzheimer Disease and Parkinson Disease, have been explored by investigating genetic pathways[13, 14]. Genetic pathways can also play an important role in drug discovery. For example, targeting a specific step in a disease pathway with the aim of identifying highly specific inhibitors can be used in drug development efforts[15]. Additionally, pathway analysis has also been shown to be useful for analyzing groups of proteins in signaling or metabolic pathways with known functions to find more effective drug targets[16].

Functional pathway analysis can be broadly classified into over-representation analysis (ORA), functional class scoring (FCS), or Pathway Topology (PT)-Based approaches[17]. In contrast to ORA or FCS, PT analysis takes into consideration structural and topological information about pathways, such as positions of genes in the pathway diagram, types of reactions, and number of reactions. This approach can be supported by knowledge within knowledge bases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG)[2], MetaCyc[18], and Reactome[19]. A potentially insightful aspect of pathway analysis includes the study of structural patterns that might be embedded within directed graphs. Studying such structural patterns could be used to identify major sub-processes that may be associated with major biological functions (e.g., regulation).

The structural analysis of genetic pathways lies at the intersection of biomedical informatics, graph theory, and data mining[20-22]. Many research efforts have been directed to the prediction and identification of pathway features of potential interest. You, *et al.* used graph substucture analysis to find biologically meaningful substructures in KEGG's metabolic pathways[22]. Cakmak and Ozsoyoglu showed that functionality patterns in metabolic networks enriched with functional annotation of enzymes could be used to discover unknown pathways in organisms[23]. Battle, *et al.* used quantitative genetic interaction measurements within a Bayesian learning framework to identify

pathways[24]. Cerami, *et al*. combined an analysis of sequence mutations with a network analysis of molecular interaction networks to identify core disease pathways in Glioblastoma[25]. Chen, *et al*. used toplogical information of graphs to find optimal set of features to answer the question whether a module of proteins forms a meaningful pathway[26]. Huang, *et al.* used feature set including graph properties, biochemical and physicochemical properties for pathway classification[21]. Many of pathway analysis studies combine graph structure information, knowledge about genes and proteins at functional and biochemical levels.

The focus of this study was on the structural pattern analysis of genetic pathways of diseases. The particular goal of the study was to identify major components that may characterize disease classes, focusing primarily on complex disorders (i.e., disorders that involve multiple genes). For each disease category, distinctive functional and structural characteristics ('fingerprints') were identified based on the training of a classification model using genetic pathways dataset.

Methods

The overall goal of this study was to develop an approach to identify unique characteristics ('fingerprints') associated with a given disease class. The process started by annotating elements within a training set of disease pathways with functional annotations. These functionally annotated pathway graphs were then structurally analyzed to learn a probability model that accounted for both the graph structure and functional annotations. This model was used in pathway classification to assess the effectiveness of learning disease characteristics.

Functional Annotation of the KEGG Pathways Dataset

KEGG pathways are stored in files formatted according to the KEGG Markup Language (KGML), used to model genetic pathways. The KGML files were parsed using the BioRuby API[27] to extract nodes and edges that composed a directed graph. Edges were annotated in the KGML files with 'relation' labels such as "activation," "phosphorylation," and "expression." Nodes that represented genes were further annotated with functional annotations using the Gene Ontology (GO), based on information extracted from the Human Protein Reference Database (HPRD)[28].

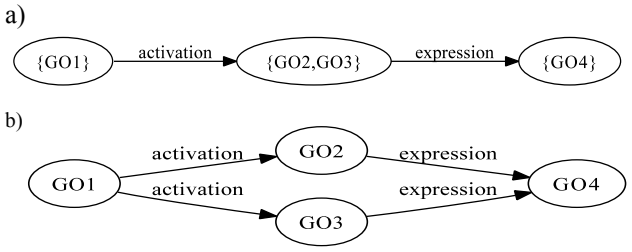

Figure 1. Node and Edge replication. A node that has more than one GO annotation in graph (a) has been replicated in graph (b). As a consequence, edges have also been replicated in (b).

Each node in a KGML file can represent more than one gene. Furthermore, each gene may match more than one GO term in HPRD annotated dataset. Thus, there can be a list of GO terms for each entry in a given pathway. Because the proposed model for classification can handle only one annotation per node or edge, a preprocessing step was developed that took nodes with more than one GO term and replicated them. Furthermore, each replicated instance of a given node carried only one GO term. Whenever a node was replicated, its incoming and outgoing edges were copied to link replicated nodes to their predecessor and successor nodes. Figure 1 illustrates this process.

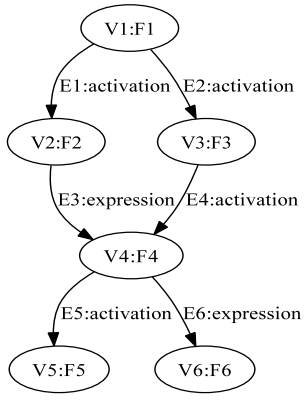

Figure 2. A labeled directed graph that represents a functionally annotated genetic pathway.

Graph Representation of Genetic Pathways

Disease pathways were modeled as labeled directed graphs where nodes represented genes and edges represented relationships between genes. An example of a labeled graph is shown in Figure 2. Node V_3 has a label F_3 and Node V_4 has label F_4. An edge (E_4) connecting this pair of nodes has the label 'activation'. In addition to labeled nodes and edges, each disease pathway was associated in KEGG with a class label categorizing the nature of the disease. Examples of pathway class labels in the dataset include: 'cancer,' 'infectious,' and 'immune.'

Mathematical Model

A particular class of diseases was assumed to have specific characteristics that make it distinct from other disease classes. The implemented model thus took into account associations between a particular disease class and pathway's structure and annotations. Every graph instance, *G*, was considered as one of many possible examples that contained characteristics of a disease class *C*. Every pair of disease class and graph *(C,G)* was assigned a probability value *P(G|C)*, which was interpreted as a quantification of the amount of characteristics of disease class *C* contained in graph G. The system then aimed to find disease class *C,* given an observed graph *G*. These relationships can be expressed using Bayes' theorem:

$$P(C \mid G) = \frac{P(C)P(G \mid C)}{P(G)} \qquad (1)$$

Then, the goal of the classifier is to search for a disease class $\hat{C}$ for which *P(C|G)* was the greatest, where

$$\hat{C} = \arg\max_{C} P(C)P(G \mid C) \qquad (2)$$

This makes the assumption that the denominator of Eq. 1 was independent of *C*, thus suggesting that finding $\hat{C}$ was the same as finding *C* so that the quantity *P(C)P(G|C)* was as large as possible.

Incorporation of Graph Substructures

The calculation of probability value *P(G|C)* needed to take into account the possible structural patterns of G that could be considered characteristics of disease class *C*. A given graph can be decomposed in many ways into subgraphs, each of which can be considered a candidate characteristic of a disease class. The decomposition of a graph into its sub-graphs was defined using the following definition for graph partitioning:

Definition 1. Partitioning *A partitioning Φ of graph G is a function Φ:E(G) → N, where E(G) is the edge set of G and N is the set of natural numbers. A subset of edges {e_1, e_2 ... e_k} is said to be in the same subgraph if and only if Φ(e_1) = Φ(e_2) ... = Φ(e_k).*

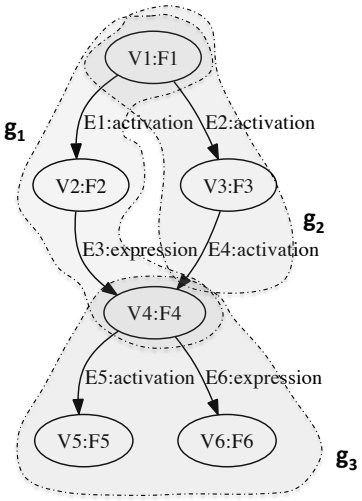

Figure 3. A partitioning of graph G into three subgraphs g1, g2 and g3

A partitioning of a given graph is a set of subgraphs that are edge-disjoint (i.e., an edge belongs only to one subgraph). This partitioning can be represented by an array of integers where positions points to edges and content indicate a subgraph to which the edge in position belongs. To illustrate this definition, consider the following example. Let *E(G)= <E_1, E_2, E_3, E_4, E_5, E_6>* be an ordered sequence of edges in a graph G. A partitioning can be represented as an integer array of length equal to the *|E(G)|*. A set of subgraphs S is created according to this partitioning. For each subgraph $g_i \in S$, edge set of g_i is $E(g_i)= \{e \mid \Phi(e)=i\}$. An example of a partitioning Φ is the sequence *<1,2,1,2,3,3>,* which means that G can be divided into three subgraphs: g_1 containing edges *{E_1, E_3}*, g_2 containing the edge *{E_2, E_4}*, and g_3 containing the edges *{E_5, E_6}*. Figure 3 shows an example of this kind of partitioning.

To incorporate structural patterns in the calculation of *P(G|C)*, graph partitioning can be introduced as a hidden variable Φ, and hence *P(G|C)* can be expanded as:

$$P(G \mid C) = \sum_{\Phi} P(\Phi, G \mid C) \qquad (3)$$

A partitioning function naturally divides a graph into a set of features that can be used for classification. Since there might be many possible partitionings (some of them might be equally probable), a sum over partitionings is used in the right hand side of Eq. 3. Let $S = \{g_1, g_2 \ldots g_n\}$ be the set of subgraphs of G according to a partitioning *Φ*. The likelihood of an arbitrary partitioning *Φ* of graph *G* given a class *C* can be expressed as:

$$P(\Phi, G \mid C) = \prod_{g \in S} P(g \mid C) \qquad (4)$$

Where, S is the set of non-overlapping subgraphs of graph G according to partitioning Φ: $S = \{g_i \mid \forall e \in E(g_i), E(g_i) \subseteq E(G), \Phi(e) = i\}$.

The product of $P(g|C)$ terms used in Eq. 4 assumes that subgraphs or features are orthogonal. The value $P(g|C)$ represents the likelihood that g is a characteristic of class C. To simplify calculation of $P(g|C)$, a subgraph g can be approximated by a set of maximal paths, A, inside g. Hence,

$$P(g \mid C) \approx \prod_{A \in g} P(A \mid C) \quad (5)$$

And then, one can combine Equations 4 and 5:

$$P(\Phi, G \mid C) = \prod_{g \in S} \prod_{A \in g} P(A \mid C) \quad (6)$$

Where, $A \in g$ denotes a maximal path A inside subgraph g. Finally, the probability of a given partitioning Φ can be calculated using

$$P(\Phi \mid G, C) = \frac{P(\Phi, G \mid C)}{\sum_j P(\Phi_j, G \mid C)} \quad (7)$$

Equations 3-7 suggest an approach to compute the conditional probability $P(G|C)$ in a tractable manner. To compute this probability, it was first required to generate partitionings. Then, the likelihood of each partitioning was calculated based on a conditional probability distribution of paths $P(A|C)$. Thus, finding paths inside subgraphs was needed to build and update $P(A|C)$. Therefore, the training procedure was based on finding paths inside subgraphs and utilizing the concept of partitioning to compute $P(G|C)$ according to Equation 3.

Model Training

The objective of probability estimation is to build the conditional distribution $P(A|C)$. This involves counting co-occurrence of pairs of path A and a class C. Since, for any given graph, there can be many different ways to decompose it into subgraphs, a single path may simultaneously belong to more than one possible subgraph. The question is *how to count the co-occurrence of the path-class pair in this case?* A possible solution is to weigh each occurrence of path-class pair by the probability of the partitioning Φ to which that path belongs. This step is called collecting fraction counts, since each occurrence of path-class pair is discounted by the probability of partitioning Φ. The idea of collecting fraction counts has been successfully applied to machine learning problems such as statistical machine translation[29].

Counting Class-Path Co-occurrences

To collect counts of path-class pairs, the probability of possible partitionings needs to be computed. In turn, computing probability of partitionings needs the conditional probability $P(A|C)$, which depends on counting co-occurrences of path-class pairs. This problem can be solved by an iterative training procedure using the Estimation Maximization (EM) algorithm[30]. The first step is seeding the partitionings: to generate a number of random partitionings (maximum number of partitionings is an adjustable parameter of the tool) for each graph. Instances of path-class pairs, (A,C), are then identified within each subgraph according to each partitioning. The counts of (A,C) pairs are used to create the conditional probability distribution $P(A|C)$. Thus, this iterative process has two phases: (1) E-Step: search for and compute the likelihood of each partitioning using Eqs. 6 and 7; and (2) M-Step: fraction counts of (A,C) pairs are collected, and better estimates of conditional probability $P(A|C)$ is produced. The number of training iterations is an adjustable parameter. For the dataset used in this study, three iterations were used. The outline of this process is shown in Algorithm 1.

At the beginning of model training, the conditional probability table $P(A|C)$ is initialized with single-edge paths. There is a minimum probability value $\varepsilon = P(A|C)$ for paths that are not discovered yet in early iterations of EM algorithm. In the E-Step, new paths are likely to be discovered when new partitionings (and probably larger subgraphs having longer paths) are explored in Hill-Climbing search for partitionings. These newly discovered paths are added to the conditional probability $P(A|C)$ when collecting fraction counts in the M-Step.

Searching for Highly Probable Partitionings

Since there is a large set of possible partitionings for a given graph instance, it was necessary for the system to limit

the search by considering only a set of best scoring partitionings. To find such a set, the search process started with a seed partitioning (with subgraphs containing only one edge) and then using Hill Climbing technique to find its neighbors of good partitionings. A set of operators was used to modify a current partitioning by changing edge membership from one subgraph to another. This step resulted in the growing of some subgraphs and shrinking of others. A priority queue was used to store partitionings ordered by their likelihood calculated by Eq. 7.

Algorithm 1. *P(A|C)* Probability Estimation

Input:
D: graph data set $\{G_1, ..., G_n\}$
N: Number of iterations
Process
1: Create seed partitionings and Initialize *P(A|C)* table with uniform probability value.
2: for $i=1:N$
E-Step
3: for each $G \in D$
4: Let C be the class label of G and let the set of graph partitionings $G.\Phi s$ = searchForPartitionings(G,C)
5: Use Eq. 6,7 to compute the likelihood of every partitioning $\Phi \in G.\Phi s$
M-Step
6: for each $G \in D$
7: for each $\Phi \in G.\Phi s$
8: for each subgraph $g \in S = \{g_i \mid \forall e \in E(g_i), E(g_i) \subseteq E(G), \Phi(e) = i\}$
9: for each maximal path $A \in$ subgraph g
10: CountTable(A,C) += $\Phi.probability$ // collecting fraction counts
11: Normalize entries of CountTable (A,C) to obtain *P(A|C)*
Output: updated *P(A|C)*, $G.\Phi^*$ //return conditional probability and best partitioning

Predicting Class Labels

Given a conditional probability model *P(A|C)* for paths and class labels as well as a prior probability distribution model *P(C)* for class labels (P(C) can be computed using frequency of each disease class in the dataset), a new graph instance was classified as follows. A search for best partitioning for the target graph was performed, using all possible candidate class labels. The evaluation of partitioning quality was measured using *P(A|C)* according to Eqs. 4-7. The Hill-climbing search for set of best partitionings was performed. The candidate class label that maximizes Eq. 2 was reported as target class label. Algorithm 2 shows how classification was performed.

Algorithm 2. Predicting a Class Label for a Testing Graph Instance

Input: Graph G, set of class labels C, paths conditional probability distribution *P(A|C)* and prior class probability distribution *P(C)*
Process
1. For each class label $\ell \in C$
2. Using the probability distribution P (A|C), Φs = searchForPartitionings(G, ℓ)
3. Compute P(ℓ)P(G|ℓ) according to Eqs. 3-6 using the set of partitionings Φs
Output: Class label ℓ with the highest P(ℓ)P(G|ℓ) value

Extracting Disease Fingerprints

Fingerprints were defined as subgraphs representing structural patterns and were extracted from the best partitionings of graph instances. These sets of subgraphs were considered key characteristics of disease classes and highlight major processes (e.g., chains of reactions) inside a disease pathway. Fingerprints were extracted from the best partitionings that had probability scores greater than a specified threshold value $(\delta > 0.1)$, which represents confidence about partitioning quality. The choice of the threshold value depends on the size of graphs (in terms of edges) and the number of graphs in the dataset. This threshold helps one to choose only highly probable partitionings for manual inspection. High threshold values tend to print low numbers of partitionings to disk files. If more partitionings need to be examined, a slightly lower threshold value can be used. To show the importance of structural patterns in identifying macro-level view of each disease pathway, maps were generated to represent joint

distributions of GO terms. The intensity in these maps reflected how often two GO terms were linked together by an edge in the graph. Since edges can have annotations for a set of basic processes such as expression or phosphorylation, a map was generated for each process type. Thus, maps were generated for expression, phosphorylation, activation, etc. The spatial patterns of these maps enabled a global view of GO terms connectivity within the complete data set. Nodes of subgraph fingerprints were mapped onto points in maps to see if nodes of subgraph fingerprints tended to cohere (found to be near each other) in the map. The maps were developed to highlight the utility of structural patterns in contrast with micro-level patterns that emerge from graph-theoretic properties such as edge degree distribution.

Experimental Settings

Datasets

Pathway diagrams were downloaded from KEGG Pathway database (December 2011). GO annotations were extracted from the Gene Ontology file of HPRD. This dataset was composed of 56 *KEGG's* disease pathways. The numbers of pathways per each class category as defined by KEGG are shown in Table 1.

Evaluation Metrics

A goal class label was defined as a specific disease category that the binary classifier should report as positive example. For example, "Cancer" could be a goal class label, in contrast to the "NonCancer" class label, which should be reported as negative example. True positives (*TP*) were defined as instances with goal class label and to which the classifier assigned goal class labels. False negatives (*FN*) were defined as instances with goal class label that were assigned non-goal class labels by the classifier. False positives (*FP*) were defined as instances with non-goal class labels to which the classifier assigned goal class labels. In this study, accuracy was measured as the geometric mean of Positive Predictive Value (*PPV*) and Sensitivity (S_n), where $PPV = \frac{TP}{(TP+FP)}$ and $S_n = \frac{TP}{(TP+FN)}$. Finally, accuracy was defined as: $A_g = \sqrt{PPV \times S_n}$.

Table 1. KEGG Disease Pathway Classes

Disease Class	Instances
Cancer	15
Infectious	22
Substance Dependence	1
Neurodegenerative	5
Immune	7
Cardiovascular	4
Metabolic	3

Experiments

Each pathway in the dataset was annotated with *GO* terms of the manually curated *HPRD* database. By excluding the Substance Dependence pathway data, which had only one instance, this dataset was used to test six binary classifiers, one for each disease class. For each disease class, a two-label dataset was generated. For instance, a cancer dataset was developed that contained pathways with labels 'cancer' and 'non-cancer'. Then, evaluation of accuracy of each binary classifier was measured on these six datasets. A 3-fold cross validation procedure was applied to each dataset. Given the small dataset, cross validation procedure was run for 30 iterations.

Results

Classification Accuracy

Average accuracy for each of the datasets is shown in Table 2 (based on 30 cross validation runs). For Metabolic, Cardiovascular, Neurodegenerative, and Substance Dependence datasets, the system was not able to classify any positive classes correctly (TP value was zero), due to the small number of instances of these classes in dataset. However, the total number of instances of these classes was 20, therefore including them, as negative examples of cancer and infectious disease, in training data was important.

Table 2. Average Classification Accuracy

Disease Class	Accuracy	PPV	S_n
Cancer	0.8	0.77	0.83
Infectious	0.67	0.6	0.75

Disease Fingerprints

As a by-product of the EM training process, the best partitioning of each pathway graph was saved to output files. Each partitioning highlighted a set of subgraphs (features) inside a pathway. Annotating nodes of pathways with

functional annotations (e.g., GO terms) yielded an abstract representation of pathways. Thus, the subgraphs identified inside each pathway could be regarded as functional sub-units. Each category of disease was assumed to have its characteristic functional units (fingerprints) inside pathways under that category. Individual GO terms could be found equally in pathway graphs of two different disease classes. Correlation tests may not be able to find a strong association between a disease class and a given GO annotation of genes in pathway graphs. However, the conditional probability distribution of paths and disease classes suggested that some paths tend to be found more frequently in a specific class of diseases and less frequently in other classes. Table 3 shows a number of paths that tend to appear in cancer pathways and those that tend to appear in non-cancer disease pathways.

Table 3. Paths Associated with Cancer/Non-Cancer

Annotated Path	Disease Class
GO:0003924-activation-GO:0004674# GO:0004713-inhibition-GO:0004713# GO:0003924-activation-GO:0030159#-GO:0030159-activation-GO:0004674# GO:0003924-activation-GO:0004674#-GO:0004674-phosphorylation-GO:0004712#	Cancer
GO:0030528-dissociation-GO:0003700# GO:0004713-phosphorylation-GO:0003700# GO:0005509-binding/association-GO:0005200#-GO:0005200-binding/association-GO:0005198# GO:0004930-activation-GO:0003924#-GO:0003924-indirect_effect-GO:0004620#	Non-Cancer

The maps shown in Figure 4(e) give macro view of the linkage of GO terms in annotated graphs and demonstrate that distribution of pairs of GO terms is sparse. In general, this view does not suggest much about the structure of graphs. Instead, they reflect the fact that although there are many edges connecting GO terms in the graph dataset, only few GO term pairs are linked more frequently than other. However, structural patterns that are uncovered in graphs can be important to learn facts about key functional components inside graphs. Figure 4(a)-(d) shows such structural patterns, which are linked to maps of basic processes of expression, phosphorylation, and activation as shown in Figure 4(e). The mapping of edges of these structural patterns into maps in Figure 4 suggests that biological meaningful patterns do not necessarily correspond to spatial patterns in maps. This might be because functional similarity is not the only reason to link two genes in a given disease pathway. Functionally dissimilar genes might be found linked in a pathway, and thus it would be expected to find dissimilar (spatially distant) GO terms to be linked in a disease fingerprint (subgraph), but found spatially scattered in the map. It should also be mentioned that the method presented here allows for the inclusion of some edges in a disease fingerprint (subgraph) even though these edges are not directly related to that disease.

Discussion

Extracting meaningful structural patterns (fingerprints) of disease categories is important for many reasons. Meaningful patterns can illustrate interactions between proteins in functional terms that would help better understanding of genetic pathways. This, in turn, can help biomedical and pharmacological researchers indentify important biological sub-processes that might take place inside cells. From a knowledge discovery perspective, identifying sets of fingerprints of disease pathways can be important for mining tasks such as discovery and classification of disease pathways. In this study, a probabilistic model was developed to identify such important substructures (disease fingerprints) in functionally annotated pathways. Identified disease fingerprints were used in classification of test set of disease pathways into disease categories.

The synergy of different sources of biological knowledge bases and computational models is important to uncover patterns of interest. Biochemical, physicochemical, graph based properties are integrated in models of analysis of large biological networks. Using network properties alone can help in studying of structure and general dynamics of networks, while looking for useful and meaningful patterns would require incorporation of knowledge sources. Functional annotations have been shown important for discovery, analysis, and classification of genetic pathways based on biological functions[23, 25, 31, 32]. In this study, GO terms were used to enrich KEGG's disease pathway graphs with functional annotations, which were essential to represent these graphs at an abstract level. The integration of knowledge sources also requires specially designed computational models to make best use of these sources. For instance, in this study, functional annotations were incorporated in a probabilistic model that took into account associations between sets of functional annotations as represented in paths and subgraphs. In contrast, using separate functional annotations as features could be less effective than expected. For instance, only small set of GO terms was identified as optimal features and was encoded in feature vectors for graph classifications of pathway diagrams[21]. Knowledge-enriched models that make use of associations between GO terms can be more effective[23, 33].

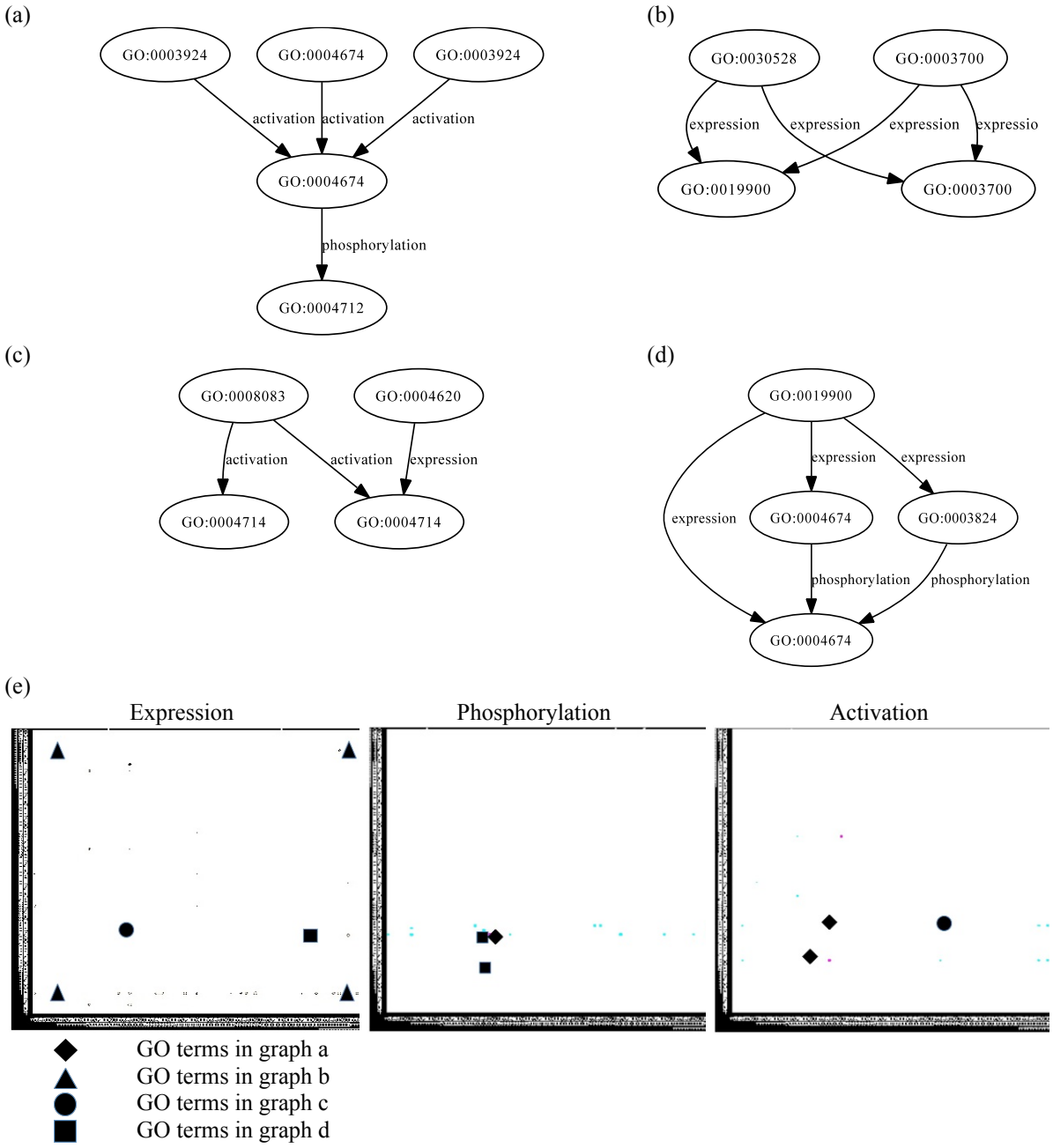

Figure 4. Disease fingerprints for cancer pathways and the mapping of their nodes onto maps that represent GO terms associations in data. Directed graphs that represent fingerprints extracted from best partitionings of cancer pathways are shown in (a)-(d). Pairs of GO terms in (a)-(d) that were part of expression, phosphorylation, and activation processes are highlighted in the maps shown in (e).

This study aimed to address a problem related to discovery of key structural patterns in graph datasets. These patterns were searched for in the training process of a graph classification model. The problem was cast as finding optimal substructure feature sets ('fingerprints'). The concept of partitioning enabled searching for features in a coherent way that is effective in avoiding irrelevant or redundant structural patterns. The proposed mathematical model and EM algorithm used the concept of partitioning to get better estimates for the conditional probability distribution for graph paths given disease classes. This idea can be similar to maximum likelihood (ML) phylogenetic analysis[34]. In a sense, ML phylogenetic analysis uses nucleotide transition probability distribution to search for more likely phylogenetic trees (which can be perceived as a hierarchy). The ML training for phylogenetic analysis produces a best scoring phylogenetic tree for a set of genes while improving parameters values for

nucleotide transition probability distribution. A similar practice was followed here: the study aimed to produce the best partitionings while improving the conditional distribution of paths given classes.

Identifying optimal feature set for graph classification is an important problem in graph data mining[35, 36]. One method for graph classification is to use graph pattern mining to generate candidate features. Then, optimal feature set for classification is identified using variety of measurements such as information gain. However, graph classification techniques that use graph pattern mining for feature selection have three major problems:

(1) *The search for features is local and sequential.* Candidate subgraph features are extracted and evaluated in isolation. The problem with this method is that features can be redundant or less informative.
(2) *The criteria used for feature selection might not be optimal.* For instance, subgraph frequency can be used as criterion for selecting features (e.g., using gSpan[37] to find candidate features). Frequent subgraphs may not necessarily be discriminative. On the other hand, some information theoretic features may not be effective. For instance, LEAP search utilized information gain to look for features. This strategy may fail in the following scenario as noted by Jin, *et al.*[35]: "When no individual pattern has high discrimination power, a group of patterns may jointly have higher discrimination power". Since LEAP search finds patterns sequentially, it is unlikely that it will find such groups of jointly high discriminative power.
(3) *It can be difficult to separate the searching and classification processes.* Separating the search for subgraph features and classification when using feature vectors can prevent the classification algorithm from using prior information about the distribution of class labels among graph instances.

To address the above problems in this study, the search for optimal feature set was integrated into the training of a probabilistic model for graph classification. The concept of partitioning and the utility of the partitioning function provided a means that naturally divided graphs into candidate features. Upon completion of model training, the best partitioning of each pathway instance provided a list of subgraphs that were considered characteristic components of a given pathway. The limited size of the design dataset and few number of instances per some disease classes made it not possible to analyze fingerprints for some disease classes. As more diseases have related processes identified in the future, it may be possible to analyze their fingerprints. Disease pathways in databases other than KEGG would be considered in a future work to overcome the limits of small dataset size. The scalability of this method to larger networks can be obtained by adjusting the maximum number of partitionings, which is an adjustable parameter of the tool as mentioned in the Methods section. By keeping smaller number of partitionings per graph instance, larger networks with increased annotations can be processed.

Conclusion

In this paper, an approach is presented for structural analysis and classification of genetic pathways of human diseases. Experiments on real data show good performance in terms of classification accuracy while identifying characteristic components inside each pathway both in training and testing examples. The highlighting of characteristic functional components ('fingerprints') inside each pathway gives justification of classification decisions and may help improve the understanding of how genetic pathways act at component level. The proposed model may be generalized to many biological networks that are modeled as annotated directed graphs.

References

1. Rual JF, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005;437(7062):1173-1178.
2. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic acids research 2010;38(suppl 1):D355.
3. Stelzl U, Worm U, Lalowski M, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005;122(6):957-968.
4. Franke L, Bakel H, Fokkens L, De Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. The American Journal of Human Genetics 2006;78(6):1011-1025.
5. Caspi R, Foerster H, Fulcher CA, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/Genome Databases. Nucleic acids research 2008;36(suppl 1):D623-D631.
6. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics 2011;12(1):56-68.
7. Rudy Y, Ackerman MJ, Bers DM, et al. Systems approach to understanding electromechanical activity in the human heart. Circulation 2008;118(11):1202-1211.

8. Karnovsky A, Weymouth T, Hull T, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics 2012;28(3):373-380.
9. Novoyatleva T, Diehl F, Van Amerongen MJ, et al. TWEAK is a positive regulator of cardiomyocyte proliferation. Cardiovascular research 2010;85(4):681.
10. Liu N, Olson EN. MicroRNA regulatory networks in cardiovascular development. Developmental cell 2010;18(4):510-525.
11. Slattery ML, Wolff RK, Curtin K, et al. Colon tumor mutations and epigenetic changes associated with genetic polymorphism: Insight into disease pathways. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 2009;660(1-2):12-21.
12. Cogswell JP, Ward J, Taylor IA, et al. Identification of miRNA changes in Alzheimer's disease brain and CSF yields putative biomarkers and insights into disease pathways. Journal of Alzheimer's disease 2008;14(1):27-41.
13. Lambert JC, Grenier-Boley B, Chouraki V, et al. Implication of the immune system in Alzheimer's disease: evidence from Genome-wide pathway analysis. Journal of Alzheimer's disease 2010;20(4):1107-1118.
14. Pan T, Kondo S, Le W, Jankovic J. The role of autophagy-lysosome pathway in neurodegeneration associated with Parkinson's disease. Brain 2008;131(8):1969.
15. Pawson T, Linding R. Network medicine. FEBS letters 2008;582(8):1266-1270.
16. Arrell D, Terzic A. Network systems biology for drug discovery. Clinical Pharmacology & Therapeutics 2010;88(1):120-125.
17. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Computational Biology 2012;8(2):e1002375.
18. Karp PD, Riley M, Paley SM, Pellegrini-Toole A. The metacyc database. Nucleic acids research 2002;30(1):59-61.
19. Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: a knowledgebase of biological pathways. Nucleic acids research 2005;33(suppl 1):D428-D432.
20. Maudsley S, Chadwick W, Wang L, Zhou Y, Martin B, Park SS. Bioinformatic approaches to metabolic pathways analysis. Methods in molecular biology (Clifton, NJ) 2011;756:99.
21. Huang T, Chen L, Cai Y-D, Chou K-C. Classification and Analysis of Regulatory Pathways Using Graph Property, Biochemical and Physicochemical Property, and Functional Property. PLoS One 2011;6(9):e25297.
22. You C, Holder L, Cook D. Substructure Analysis of Metabolic Pathways by Graph-Based Relational Learning. Biomedical Data and Applications 2009:237-261.
23. Cakmak A, Ozsoyoglu G. Mining biological networks for unknown pathways. Bioinformatics 2007;23(20):2775.
24. Battle A, Jonikas MC, Walter P, Weissman JS, Koller D. Automated identification of pathways from quantitative genetic interaction data. Molecular Systems Biology 2010;6(1).
25. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS One 2010;5(2):e8918.
26. Chen L, Huang T, Shi XH, Cai YD, Chou KC. Analysis of protein pathway networks using hybrid properties. Molecules 2010;15(11):8177-8192.
27. Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T. BioRuby: Bioinformatics software for the Ruby programming language. Bioinformatics 2010;26(20):2617.
28. Prasad TSK, Goel R, Kandasamy K, et al. Human protein reference database - 2009 update. Nucleic acids research 2009;37(suppl 1):D767-D772.
29. Brown PF, Pietra VJD, Pietra SAD, Mercer RL. The mathematics of statistical machine translation: Parameter estimation. Computational linguistics 1993;19(2):263-311.
30. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 1977:1-38.
31. Hu H, Yan X, Huang Y, Han J, Zhou XJ. Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics 2005;21(suppl 1):i213.
32. Liu G, Wong L, Chua HN. Complex discovery from weighted PPI networks. Bioinformatics 2009;25(15):1891.
33. Bebek G, Yang J. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. BMC bioinformatics 2007;8(1):335.
34. Felsenstein J. Inferring phylogenies: Sinauer Associates; 2004.
35. Jin N, Young C, Wang W. Graph classification based on pattern co-occurrence. 2009.
36. Ranu S, Singh AK. Graphsig: A scalable approach to mining significant subgraphs in large graph databases. 2009.
37. Yan X, Han J. gSpan: Graph-based substructure pattern mining. 2002.