Journal of Biomedical Informatics 47 (2014) 178–191

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Structural network analysis of biological networks for assessment of potential disease model organisms

Ahmed Ragab Nabhan [a,c,d], Indra Neil Sarkar [a,b,c,*]

[a] Center for Clinical & Translational Science, University of Vermont, Burlington, VT, USA
[b] Department of Microbiology & Molecular Genetics, University of Vermont, Burlington, VT, USA
[c] Department of Computer Science, University of Vermont, Burlington, VT, USA
[d] Faculty of Computers & Information, Fayoum University, Al Fayoum, Egypt

ARTICLE INFO

Article history:
Received 13 January 2013
Accepted 21 October 2013
Available online 5 November 2013

Keywords:
Disease pathway mining
Translational bioinformatics
Structural pattern analysis
Interaction networks

ABSTRACT

Model organisms provide opportunities to design research experiments focused on disease-related processes (e.g., using genetically engineered populations that produce phenotypes of interest). For some diseases, there may be non-obvious model organisms that can help in the study of underlying disease factors. In this study, an approach is presented that leverages knowledge about human diseases and associated biological interactions networks to identify potential model organisms for a given disease category. The approach starts with the identification of functional and interaction patterns of diseases within genetic pathways. Next, these characteristic patterns are matched to interaction networks of candidate model organisms to identify similar subsystems that have characteristic patterns for diseases of interest. The quality of a candidate model organism is then determined by the degree to which the identified subsystems match genetic pathways from validated knowledge. The results of this study suggest that non-obvious model organisms may be identified through the proposed approach.

1. Introduction

Complex diseases stem from an interplay of genetic and environmental factors. At the genetic level, these diseases are often associated with the dysfunction of more than one gene. This necessitates the study of complex diseases at a systems level, which includes the modeling of cellular processes that underlie an observed disorder and may involve both sequential and simultaneous molecular interactions between many agents (e.g., genes and chemical compounds). This highlights the importance of curating molecular interaction networks (e.g., gene/protein interaction networks, metabolic networks, and genetic pathways). Data resources that catalogue these networks are increasing both in terms of the number and size of networks as well as their coverage of organisms. Environmental factors, on the other hand, complicate the study of human diseases, since it is difficult to create a controlled environment that enables scientists to study environmental effects on disease development. Hence, model organisms offer opportunities for detailed study of features associated with complex diseases, because these organisms may be genetically engineered to produce desired phenotypes (e.g., associated with a particular disease of interest) and can be studied more easily in a controlled environment.

Model organisms play a vital role in advancing knowledge about disease processes. The sophisticated genetics of human diseases makes it important to study model organisms to uncover underlying mechanisms of diseases. Model organisms may not necessarily be closely related to humans from an evolutionary perspective. For instance, yeast are regularly used to model disease states [1]. Comparison of different phenotypes that arise from a conserved set of genes can be important for exploring model organisms for specific human disorders or diseases [2,3]. Analysis of model organism microarray data may also help identify those that have disease-related genes differentially expressed [2].

The house mouse (*Mus musculus*) has been a typical model organism in the study of human disease processes [4], as well as complex traits and social behavior [5]. Mice have also been genetically engineered to provide models for studying cancer and immune diseases [6,7]. However, mice may not always be suitable for the study of all categories of disease. In a recent study of 'phenologs' (phenotypes that are equivalent across organisms), McGary et al. suggested a worm model (*Caenorhabditis elegans*) for breast cancer, a mouse model for autism, a plant model (*Arabidopsis thaliana*) for Waardenburg syndrome, and a yeast model (*Saccharomyces cerevisiae*) for angiogenesis disorders [3]. Thus, there may be many potential choices for a suitable model organism relative to the spectrum of phenomena associated with disease. An

* Corresponding author. Address: Center for Clinical and Translational Science, University of Vermont, 89 Beaumont Avenue, Given Courtyard, N309, Burlington, VT 05405, USA. Fax: +1 802 656 4589.
E-mail address: neil.sarkar@uvm.edu (I.N. Sarkar).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.10.011

empirical approach may therefore facilitate the identification of organism(s) that might provide insights to human diseases.

Evaluation of candidate model organisms might be measured by the degree to which gene/protein interaction networks include pathways that are structurally and functionally similar to human disease-related biological processes. To this end, prediction of pathways in candidate model organisms that are similar to disease-related pathways in humans can be effective in evaluating model organisms. Pathway prediction can be performed by a variety of techniques. A widely used technique involves mining gene or protein interaction networks to extract dense subgraphs (highly connected components within the network) and then calculating the statistical significance of the discovered subgraphs [8]. Statistically significant subgraphs are then cast as predicted pathways. Tian et al. developed a method to discover statistically significant pathways from gene expression data [9]. Bebek and Yang annotated gene networks with GO annotations and developed the Path-Finder method to predict novel pathways [10]. Cakmak and Ozsoyoglu developed a method that used frequent functional patterns in a known pathway to find organism-specific versions of that pathway in the gene networks [11]. Finally, Senf and Chen developed a hidden Markov model-based method to identify genes participating in genetic pathways [12].

The present study proposes a computational method that attempts to provide a quantitative measure of how well a candidate model organism might be suited for the study of a given disease type. The proposed quantitative measure is based on the proportion of correctly predicted genetic pathways that can be identified in interaction networks for a given organism. The proposed approach makes use of three types of knowledge resources: (1) Kyoto Encyclopedia of Gene and Genomes (KEGG) [13] pathway database, (2) The Biological General Repository for Interaction Datasets (BioGRID) and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) gene/protein interaction databases [14], and (3) Gene Ontology (GO) [15] annotations that have been applied to genes or proteins in curated databases. The main premise of this work was to leverage a machine learning method to extract significant functional and structural patterns, or 'fingerprints,' [16] from functionally annotated KEGG disease pathways and match these patterns to functionally annotated gene/protein interaction networks in major databases (e.g., BioGRID) as well as meta-databases (e.g., STRING). Depending on an organism's interaction network coverage of structural patterns for a given disease, it can be ranked in terms of model organism suitability for that disease. Through the use of a statistical model, this study was able to quantify the dependency of functional structural patterns in pathways and disease categories for 14 organisms. It was assumed that some species may be a better suitable model for one disease category and thus less suitable for studying other diseases. This assumption was motivated by the McGary et al. study, where a range of model species were suggested for complex diseases [3]. The promising results suggest that the described approach may be used to determine the potential for a given organism to serve as a model for the study of a particular disease.

2. Materials and methods

In this section, the five phases of the developed approach are described: (1) annotation of gene/protein nodes in pathway graphs with molecular function annotations, (2) learning disease fingerprints within annotated pathways, (3) functional annotation and indexing of gene/protein interaction networks, (4) prediction of novel subsystems within gene/protein interaction networks using learned fingerprints, and (5) scoring discovered subsystems using reference pathways. Fig. 1 provides an overview of the approach.

2.1. Functional annotation of KEGG pathways

KEGG genetic pathways are modeled as directed graphs with a node set (V) representing biochemical entities such as genes, chemical compounds, and protein complexes and an edge set (E) representing interaction relations between entities such as general process type (e.g., a gene expression [GErel] or protein interaction [PPrel] relation) and specific relation types (e.g., activation, expression, and inhibition). For this study, only gene/protein nodes were considered. To increase the generalization capability, gene nodes were enriched with molecular function annotations as defined in Gene Ontology (GO) [15]. These GO annotations were imported from Human Protein Reference Database (HPRD) [17] and overlaid on gene/protein nodes of pathway graphs. Gene/protein nodes without a match to HPRD GO annotations were assigned a default 'NULL' annotation. Nodes could be associated with multiple GO term annotations and edges could also have multiple labels. Thus, for each graph there was a shift of focus from "what gene/protein is in a given node?" to "what function does the node perform in a system that models a biological process?" With knowledge-enriched annotations of genes/proteins, pathways were represented at a functional level. Subsequently, functional structural patterns in these pathways graphs could be matched to sub-networks of large interaction networks with functionally annotated nodes. In this study, the KEGG disease pathways dataset contained 63 disease pathways across seven human disease classes. KEGG disease pathways cover many biological processes related to genetic information processing, metabolism, and cellular processes. However, this study did not focus on a particular pathway category such as metabolic pathways and cellular processes. Each graph instance in this design set was associated with a class label from the seven disease classes in KEGG.

2.2. Learning disease fingerprints

The objective of the second module of the proposed method was to identify characteristic biological functionality patterns, termed "fingerprints," in annotated disease pathways. A mathematical model and an algorithm were designed to accomplish this task. A disease fingerprint was defined as a subgraph within a GO annotated disease pathway. Fingerprints were assumed to represent functional sub-processes that could be characteristic of a disease class such as immune, infectious, or neurodegenerative disease. Graphs in the design dataset were assumed to be independent and identically distributed (*iid*) data observed from an unknown probability distribution $P(G)$. The *iid* data assumption was made to facilitate statistical inference and to make decision about properties (e.g., class label) of a graph instance independent of other graph instances in the dataset. For a given GO-annotated pathway graph, there can be a large number of possible GO functionality subgraph patterns, which will be called "subgraph patterns" hereafter. A mathematical model was proposed to allow for scoring of subgraph patterns. High scoring patterns were output from the model as disease fingerprints.

Mining of key subgraph patterns in the dataset was performed so that a subgraph pattern is evaluated within a context of its neighboring patterns in a graph. To formalize the idea of neighbor context, a utility function termed "graph partitioning function" was used to decompose a graph into a set of subgraphs. A partitioning function π: $E(G) \rightarrow Z$ assigned an integer to every edge e of graph edge set $E(G)$ such that edges with the same integer formed a subgraph. The set of subgraphs $H\pi$ that were highlighted by a specific partitioning function (π) was defined as $H\pi = \{g_i | \forall e \in E(g_i),\ E(g_i) \subseteq E(G),\ \pi(e) = i\}$. Fig. 2 illustrates the concept of partitioning.

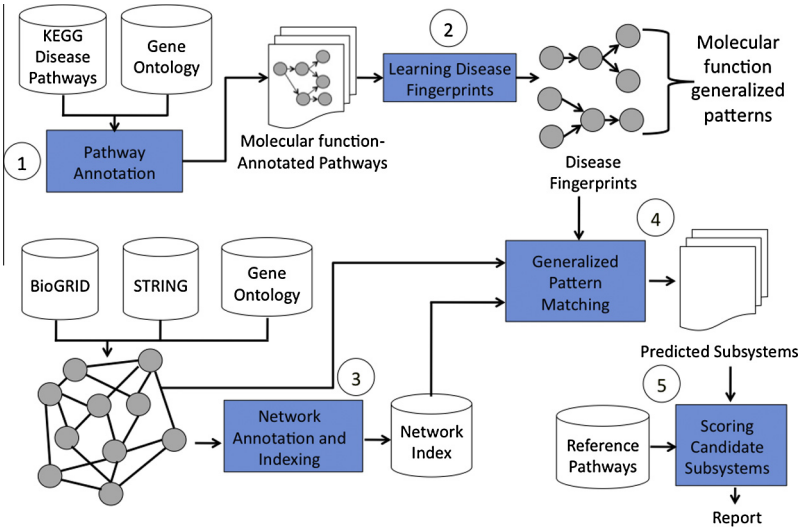


Fig. 1. Overview of the five components of the method developed in this study.

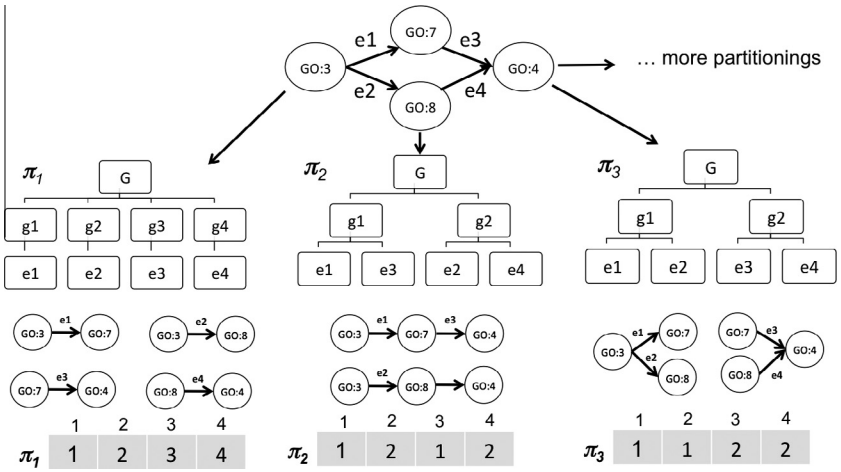


Fig. 2. An example graph is partitioned into smaller subgraphs using partitioning functions π_1, π_2 and π_3. The vector representation of each partitioning is presented under each of the three example partitionings. For instance, partitioning π_3 assigns edges 1, 2 to subgraph 1 and edges 3, 4 to subgraph 2. Additional possible partitionings are not shown.

2.2.1. Scoring of subgraph patterns within pathways

Typically, there exists a large space of possible partitionings for a given graph. Searching for the most likely partitionings in the dataset leads to the identification of key subgraph patterns (fingerprints). Searching for best graph partitionings can be better than searching for individual subgraph patterns (e.g., as in frequent pattern mining techniques [18]). This is because a graph partitioning hypothetically decomposes a system (represented by a graph) into a set of components (subgraphs), and partitioning quality reflects how good is a partitioning in identifying key components of that system (pathway in this case).

The search for best partitionings therefore required a scoring function that could be used to assign a high score to a partitioning that highlights the most likely patterns. For a pathway graph (G) of a disease class C and a partitioning π, $P(G,\pi|C)$ was defined as the probability of observing a pathway graph G and a partitioning π given a disease class C. The value of $P(G,\pi|C)$ depended on how good that partitioning highlighted key subgraph patterns. Recall that $H\pi$ was defined as the set of subgraphs according to a partitioning function π of graph G:

$$H\pi = \{g_i | \forall e \in E(g_i),\ E(g_i) \subseteq E(G),\ \pi(e) = i\}$$

The graph partitioning probability $P(G,\pi|C)$ was then computed as a function of the set of subgraphs $g \in H\pi$:

$$P(G,\pi|C) = P(g_1, g_2, \ldots, g_n|C) \quad (1)$$

where $g_1, g_2, \ldots, g_n \in H\pi$. Assuming subgraphs resulting from a partitioning function were conditionally independent, $P(G,\pi|C)$ was written as

$$P(G,\pi|C) = \prod_{g \in H\pi} P(g|C) \quad (2)$$

The probability $P(g|C)$ represented the degree to which a subgraph g was a fingerprint of a disease class C. For the purpose of probability estimation, counting the number of instances of a given subgraph in partitionings of all graphs in a direct way was deemed impractical. This was because deciding whether two subgraphs were the same would require a test of subgraph isomorphism [19]. An indirect method was thus used to approximate subgraph matching by representing each subgraph with a set of maximal paths connecting its nodes. A maximal path was defined as a path that could not be extended by adding nodes to either end. The probability $P(g|C)$ could then be expressed in terms of probabilities

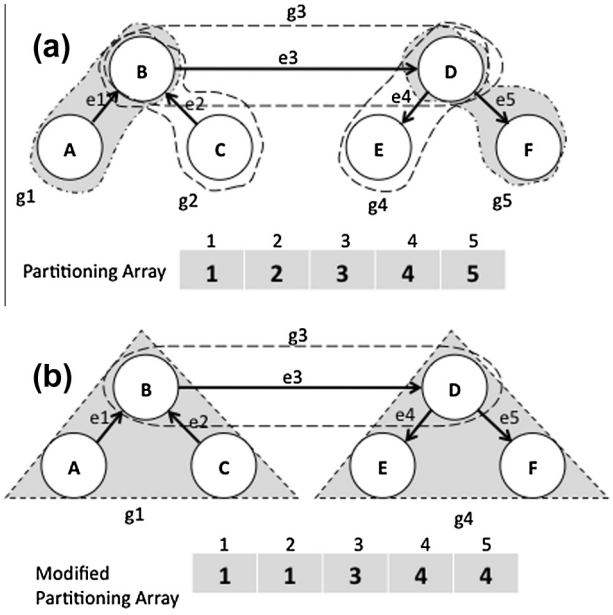


Fig. 3. Generating new partitions from an existing graph. (a) There are five subgraphs in this partitioning. Edges e_1 and e_2 belong to different subgraphs but share node B. Therefore, the partitioning vector is modified to change membership of e_2 to belong to subgraph containing e_1. Similarly, edge e_3 was assigned to subgraph g_4. The net effect is in (b) with a partitioning with 3 subgraphs.

of maximal paths given a class C. GO-annotated maximal paths inside the subgraphs were used to approximate representation of subgraphs, and thus avoid subgraph isomorphism test. Each maximal path represented a sequence of GO annotations of nodes that lay in that maximal path. In the case where a node had more than one GO annotation, multiple maximal paths were generated so that each maximal path had only one GO annotation per node. Then, $P(g|C)$ was calculated approximately as:

$$P(g|C) \approx \prod_{a \in g} P(a|C) \tag{3}$$

where a denotes a GO-annotated maximal path that connected a subset of nodes inside subgraph g. Using Eqs. (2) and (3), the likelihood of a partitioning and a graph instance given a disease class label was written as

$$P(G, \pi|C) = \prod_{g \in H\pi} \prod_{a \in g} P(a|C) \tag{4}$$

Thus, Eq. (4) represented a scoring function that was used in the search for best partitionings that highlighted disease fingerprints within pathway graphs. The problem was then that the probability distribution of maximal paths $P(a|C)$ did not exist *a priori* and needed to be estimated while searching for best partitionings. To solve this problem, an iterative training algorithm was used (described in the next section).

2.2.2. Parameter estimation

The proposed model had a set of parameters $\theta = \{P(a|C)\}$ composing entries of the conditional probability table of maximal paths. The parameter set θ needed to be estimated in order to score graph partitionings. The Expectation Maximization (EM) [20] algorithm was used to estimate model parameters according to Eqs. (2)–(4) while identifying the set of best partitionings for each graph in the pathway dataset. Initially, a set of random partitionings was generated and maximal paths within these partitionings were collected and an initial distribution for $P(a|C)$ was created. The parameter estimation process for this study had two basic steps. The first step was to search for highly scoring partitionings using the most recent probability table $P(a|C)$ obtained in the previous iteration of EM. Then, counts of maximal paths were collected from subgraphs of the set of best partitionings obtained. Collected counts were then normalized to produce a conditional probability model. During searching and scoring of partitionings, a small probability value was used as a value of $P(a|C)$ in the case where a maximal path had not been added yet to the probability table. The EM algorithm was run for four iterations in this study. Additional details of the mathematical model and EM parameter estimation procedure are presented in Appendix A.

The EM algorithm had two outputs: the conditional probability table $P(a|C)$ and the set of best partitionings of each pathway in the dataset. Disease fingerprints were extracted from best partitionings of pathways. Using the model described above, the search for disease fingerprints not only depended on an individual score of a subgraph (according to Eq. (3)), but also based on the contributions of other subgraphs in the quality of a graph partitionings (according to Eq. (4)). To test the model, a graph classifier was built to classify pathways using probability table $P(a|C)$ that was estimated during EM run. This classification task served as benchmarking of the proposed model.

2.2.2.1. Benchmarking of the fingerprint mining method. The efficacy of the structural pattern analysis method was demonstrated by implementing a graph classifier for disease pathways that utilized the conditional probability model estimated during model training. Given a test set of graphs, the task of the classifier was to assign the most likely disease class to each graph in a test set.

2.2.2.2. Classification of pathways. This classification task was modeled mathematically by finding the value for C that maximized $P(C|G)$, which represented the probability that C is a disease class of pathway G. Using Bayes' theorem:

$$P(C|G) = \frac{P(C)P(G|C)}{P(G)} \tag{5}$$

where $P(C)$ quantified *a priori* knowledge about class label distribution, $P(G|C)$ was defined as the conditional probability of observing graph G given that its class label was C, and $P(G)$ was the probability distribution of graphs. The choice of class label did not depend on $P(G)$. Therefore, Eq. (5) was expressed as

$$P(C|G) \propto P(C)P(G|C) \tag{6}$$

Modeling $P(G|C)$ directly would have required counting number of instances of a graph G. This approach had a practical challenge: Because each pathway was represented only once in the dataset, $P(G|C)$ would have followed a uniform distribution with probability equal to 1/(number of pathways of class C), and that would not have helped the statistical inference process. An alternative approach to model $P(G|C)$ that was used in this study was to incorporate the subgraph patterns in G according to the set of partitioning. Subgraphs tended to be more frequent in the dataset than their super graphs. Since one cannot be sure about which partitioning is the best, $P(G|C)$ was expressed in this study as the sum of best partitionings for graph G,

$$P(G|C) = \sum_{\pi} \prod_{g \in H\pi} \prod_{a \in g} P(a|C) \tag{7}$$

Hence, having a prior distribution $P(C)$ and a conditional probability $P(G|C)$ that was calculated using partitionings and maximal paths conditional probability distribution, the classification problem was to find a class label C^* that maximized Eq. (6):

$$C^* = \underset{C}{\operatorname{argmax}}\, P(C)P(G|C) \quad (8)$$

During classification process, the search for a set of best partitionings was performed for each test graph instance (in the same way it was performed during probability estimation). The classification process started with setting a hypothesized class label C^0 for a test graph. Then, a search for the best partitioning set started with class label of test graph fixed to C^0. Eq. (6) was used to evaluate $P(C^0|G)$. Then, another class label was used as a value for C^0, and a new set of partitionings was searched for and Eq. (6) used to calculate $P(C^0|G)$. The class label that achieved the highest score was reported as classifier output.

After benchmarking the graph structural pattern analysis method, the next module used the identified GO functionality patterns to predict subsystems in the GO-annotated interaction networks for a set of 14 species. This pattern matching module had two components, which are described in the following two subsections.

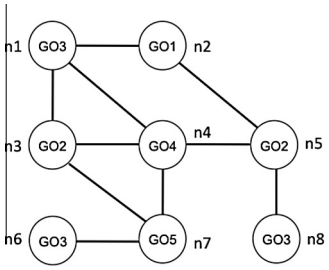
2.3. Functional annotation and indexing of gene/protein interaction networks

For each species, a network of genetic and protein interactions was constructed by importing interactions from two sources: BioGRID [14] and STRING [21]. BioGRID data contains curated interactions from high throughput datasets and individual focused studies. In this study, only interactions within the same species were included. For some species analyzed in this study, the number of interactions was limited in BioGRID. To increase coverage of a species' interaction network, more interactions were imported from STRING database (version 9.0). STRING provides information about experimental and predicted interactions. Seven sources of information about a given interaction are used in STRING, including: genome context methods, gene co-expression, text mining, as well as associations known from other database resources such as BioCyc [22] and PDB [23]. An interaction in STRING database has a combined score that is computed using evidence scores from each data source. In this study, for data imported from STRING database, only interactions with combined score greater than or equal to 70% confidence were used in the construction of networks. Since fingerprints consisted of only GO terms (i.e., not gene/protein names) interaction networks of each species were GO-annotated in order to be suitable to match disease fingerprints learned from GO-annotated disease pathways. Nodes of interaction networks were annotated with molecular function annotations from the AmiGO Gene Ontology database [24].

In this study, an interaction network of a given species could have had as many as 12,000 nodes (genes/proteins) and as many as 50,000 edges (interactions). Network indices were created for these large networks to enable efficient sub-network searches.

An index of an interaction network was built by generating a hash table with keys composed of ordered pairs of GO terms with first component being the node identifier of the node being indexed and second component denoting one of its neighbors. Values in the index table are identifiers of nodes with label equal to the first component of the ordered pair key. A value of a given key can be a single node identifier or a set of node identifiers. The index table was constructed by traversing every node in a given interaction network and examining its neighboring nodes.

Fig. 4 shows an example of GO-annotated interaction network and its index. In this example, suppose node *n1* is to be indexed. Its neighbor nodes are {*n2*,*n3*,*n4*}. For the pair (*n1*,*n2*) the corresponding annotation pair is (*GO3*,*GO1*). A key of (*GO3*,*GO1*) is inserted into the index with value {*n1*}. Similarly, the key (*GO3*,*GO2*) is inserted with value {*n1*}. In case a key already exists,



Index Table	
Key	Value
(GO1,GO2)	{n2}
(GO1,GO3)	{n2}
(GO2,GO1)	{n5}
(GO2,GO3)	{n3,n5}
(GO2,GO4)	{n3,n5}
(GO2,GO5)	{n3}
(GO3,GO1)	{n1}
(GO3,GO2)	{n1,n8}
(GO3,GO4)	{n1}
(GO3,GO5)	{n6}
(GO4,GO2)	{n4}
(GO4,GO3)	{n4}
(GO4,GO5)	{n4}
(GO5,GO2)	{n7}
...	

Fig. 4. An example interaction network and an index with keys of GO annotations.

values are appended to ones that already exist. For instance, when indexing node *n8* that is annotated with *GO3*, a key (*GO3*,*GO2*) already exists in the table with value being {*n1*}. Therefore, the value set is updated by adding element *n8* and the final key:value pair will be (*GO3*,*GO2*):{*n1*,*n8*}.

2.4. Predicting novel subsystems using disease fingerprints

Disease fingerprints were identified using the method described in Section 2.2 were matched to the GO annotated interaction networks (with interactions imported from BioGRID and STRING databases) using a similarity search algorithm. This algorithm used a network index to find subnetworks that matched an input disease fingerprint. Given a query subgraph and using the network index, the algorithm went through three steps.

In the first step, an initial set of matched node identifiers (the "candidate matching set") was retrieved for every node in the query subgraph. This was performed by using GO terms of nodes of each edge in the fingerprint subgraph to search the index. The following is an example to illustrate the pattern matching process (see Fig. 5). Let v be a node in a query subgraph. For simplicity of demonstration, presume that each node has only one GO annotation. For every node u with an edge leading to v, an ordered pair of GO terms u_t and v_t was used as a key to lookup the network index. As a result, sets of node identifiers values of the corresponding key were retrieved from the table. For example, as shown in Fig. 5c, three sets of network node identifiers that matched query node *v3* (one set per neighbor). The first set resulted from the edge (*v3*,*v1*), with a key consisting of (*GO3*,*GO1*). By looking the value up in the index table, the retrieved value was the set {*n1*}. The second candidate set for query node *v3* resulted from the edge (*v3*,*v2*), with a key consisting of (*GO3*,*GO2*). By looking this value up in the index table, the retrieved value of this key was the set {*n1*,*n8*}. Similarly, the third candidate set for query node *v3* resulted from the edge (*v3*,*v4*), with a key consisting of (*GO3*,*GO4*), with {*n1*} as third candidate set for query node *v3* (see Fig. 5c). The process was repeated for every node in the query subgraph.

The second step was to examine candidate node identifier sets for each query node and to check topological constraints. A member in the candidate node set conforms to topological constraints if it has link to a member of other candidate node sets of neighbor nodes. Topological constraints were checked first by performing set intersection operation of all candidates sets of a given query node. For example, the final candidate set for query node *v3* was {*n1*,*n8*} ∩ {*n1*} ∩ {*n1*} = {*n1*}. If the set intersection operation

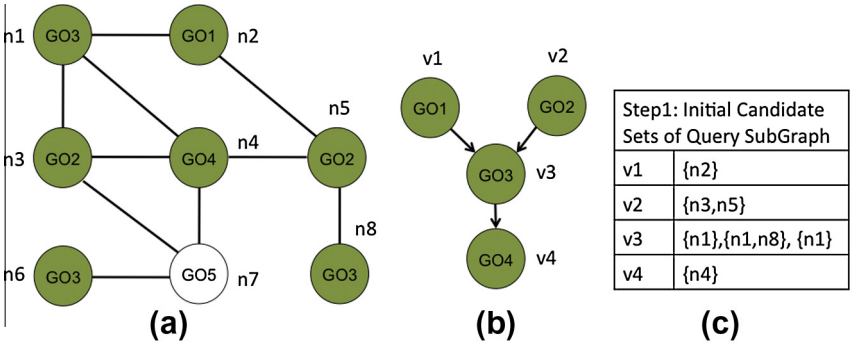


Fig. 5. The process of matching a query subgraph (GO-annotated nodes) (b) to an interaction network (a). The three steps process start with generating initial candidate set of network nodes that match the GO terms of query subgraph nodes (c). The second step ([d] and [e]) refines candidate sets by removing network nodes that do not meet topological constraints. The last step is to generate an output subnetwork as answer to a query subgraph (f).

returned empty set, then it would mean failure to match the query subgraph to any subnetwork in the interaction network, and hence the search was stopped. Node identifiers in the candidate set were then removed if they did not have any links to any node in candidate sets of other neighboring query nodes. For example, the node identifier *n5* in the candidate sets of *v2* (see Fig. 5d–e) was removed from that candidate set, because it was not connected to any item from candidate set of *v3* (*n5* was supposed to be connected to n1 according to the query subgraph structure, but in the interaction network there was no link between node *n1* and node *n5*). This step was repeated until all network node identifiers in query subgraph candidate sets satisfied topological constraints.

The third and final step was the generation of a set of subnetworks from candidate nodes sets of every query subgraph node. If there was only one node identifier for each candidate sets of query nodes, then it meant there was only one subnetwork that matched the input query subgraph. Otherwise, multiple subnetworks were returned as a matched set of the query subgraph. Details of the subgraph matching method are provided in Algorithm 1 of Appendix B. The output of this algorithm was a set of subnetworks that served as candidate subsystems that partially or completely matched known pathways available in literature.

2.5. Scoring candidate subsystems

For each disease category, fingerprints were used to find subsystems in the interaction network for each of the 14 species. To evaluate these candidate subsystems, a set of reference pathways was used to determine the degree of matching between predicted subsystems and known pathways. A candidate subsystem was considered as being predicted correctly if 70% or more of its genes/proteins were found in a known pathway in a reference dataset. The Wikipathways database [25] was used as reference dataset. As recommended on the WikiPathways download page, only the analysis collection pathways were used for evaluation. *Schizosaccharomyces pombe*, *Escherichia coli* and *Sus Scrofa* had no WikiPathways analysis collection data. Also, since the pathways of *Saccharomyces c. S288c* and *Arabidopsis thaliana* in WikiPathways data were mainly metabolic pathways, they were not used to evaluate the predicted pathways. Predicted pathways of *Escherichia coli* and *Saccharomyces c. S288c* were matched to reference pathways from BioCyc. Reference pathways for *Arabidopsis thaliana* were downloaded from AraPath database [26]. A further detailed evaluation for each species was reported for each disease of cancer and infectious disease classes in the design set.

3. Results

Evaluation of the developed approach was done in two steps. The first step was to measure the performance of the proposed mathematical model for structural pattern analysis as a function of the accuracy of a graph classifier. The second step was to evaluate the predicted subsystems that were discovered by the subgraph matching algorithm using a set of fingerprints for each disease class, and then comparing the discovered subsystems to known pathways published in the literature.

3.1. Datasets

The experiments were performed on disease pathways downloaded from KEGG pathway database (in September 2012). The

Table 1
KEGG disease pathway categories.

Disease category	Number of instances
Cancer	17
Infectious disease	22
Substance dependence	5
Neurodegenerative disease	5
Immune disease	7
Cardiovascular disease	4
Metabolic disease	3

KEGG disease pathways consisted of 63 pathways distributed over seven disease classes. This dataset is summarized in Table 1. The gene/protein nodes of the pathway dataset were annotated with GO molecular function terms imported from HPRD database. Interaction networks for 14 species were downloaded from the BioGRID and STRING databases (in October 2012). All networks were annotated with molecular function annotations from AmiGO database. The GO molecular function hierarchy included a total of 10,286 GO concepts (as of July 2012). To determine the overall accuracy of the approach presented here, the candidate subsystems identified in the 14 interaction networks were compared to published pathways in WikiPathways and BioCyc databases.

3.2. Benchmarking of structural pattern analysis model

Given the set of 63 disease pathways analyzed for this study from KEGG, two binary classifiers were developed: (1) a cancer classifier and (2) an infectious diseases classifier. Cancer and infectious diseases had the largest number of instances in the design dataset (17 cancer pathways and 22 infectious diseases pathways, respectively). Two modified datasets were created: (1) a cancer dataset where graph instances were labeled as either associated with cancer (positive case) or not associated with cancer (negative case; for this cancer classifier dataset, all non-cancer pathways such as infectious diseases, immune diseases, and neurodegenerative pathways were labeled negative); and, (2) an infectious disease dataset where graph instances were labeled as either associated with infectious disease (positive case) or not associated with infectious disease (negative case). A threefold cross validation experiment was performed. The results of classification performance in terms of the geometric average of sensitivity and specificity are shown in Table 2. An overall accuracy of 86% was achieved.

3.3. Assessment of organisms as molecular models

Assessment of organisms as molecular models was performed by matching disease fingerprints identified in disease pathways to interaction networks for 14 organisms to find candidate subsystems. Evaluation results of predicted candidate subsystems for the 14 species analyzed in this study are shown in Table 3, including the proportions of known reference pathways that were recovered by the pathway prediction method. For instance, 61% of *Bos taurus* pathways in Wikipathways were recovered. Table 3 contains the number of interactions imported from BioGRID and STRING databases. As shown in Table 3, interaction networks of *Bos taurus*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, and *Escherichia coli* achieved the top five correctly predicted pathways among the species included in this study. The number of individual and summative interactions shown in Table 3 demonstrates the impact of importing data from STRING database with regard to size of interaction network for the top five species in terms of proportion of predicted subsystems nearly matching reference pathways dataset. Some species had no data in the reference set of pathways imported from WikiPathways. In particular, *Schizosaccharomyces pombe* and *Sus scrofa* had predicted subsystems that could not be evaluated. For STRING data, zero imported interactions means that the specified threshold of evidence score was not reached or there were already enough interactions from BioGRID (e.g., *Saccharomyces c. S288c* has 234,870 BioGRID interactions and thus no additional STRING interactions were imported). *Sus scrofa* did not have any reference pathways in WikiPathways, so no prediction accuracy could be reported.

Tables 4 and 5 show detailed performance of each species with respect to individual cancer and infectious diseases. Each column in Tables 4 and 5 shows the proportion of correctly predicted pathways for each of the 14 species analyzed in this study based on matching fingerprints between disease category specific and species interaction networks. The numbers of correctly predicted pathways per species were normalized to give proportions such that each species covered a set of fingerprints for a disease. As examples of correctly predicted pathways using cancer disease fingerprints, the proposed method successfully recovered 11 out of 16 genes in the androgen signaling pathway (PW: 0000564), five out of six genes of the altered canonical *Wnt* signaling pathway (PW: 0000599) and five out of six genes in tamoxifen pharmacodynamics pathway (PW: 0000839) from the published Rat Genome Database (RGD) [27].

Table 2
Average classification accuracy.

Disease category	Average specificity	Average sensitivity	Overall average accuracy
Cancer	0.75	1.0	0.87
Infectious disease	0.86	0.82	0.84

4. Discussion

In silico identification of potential model organisms may be a cost effective first step in the study of human diseases. By annotating genetic pathways with GO terms, subgraph patterns in genetic pathways can acquire greater generalization capability. This generalization allows for matching with an organism's interaction network that was also annotated using GO terms. The degree to which an interaction network of a given model organism covered subgraph patterns of disease pathways was hypothesized to be a measure of the suitability of this model organism to study biological processes related to human diseases. A significant proportion of the interactions (genetic and physical) used in network construction were predicted interactions (e.g., inferred by genome context methods or text mining). This allowed for the evaluation of organisms as potential disease models even with limited curated interaction data.

4.1. Main findings

The statistics in Tables 3–5 show the range of disease model suitability for the 14 analyzed organisms in terms of pathways prediction accuracy. The interaction networks of *Arabidopsis thaliana* (mouse-ear cress; a plant) and *Escherichia coli* (a bacterium) performed better than those of *Gallus gallus* (chicken), *Canis lupus familiaris* (dog), or *Bos taurus* (cow) in predicting pathways using disease fingerprints of colorectal as well as thyroid cancer (see Table 4). Additionally, interaction networks of *Sacchromyces cerevisiae* (Baker's yeast) performed better than *Mus musculus* (mouse) or *Rattus norvegicus* (rat) in predicting pathways using Eppstein-Barr virus disease fingerprints (see Table 5). These types of findings are supported by McGary et al., where organisms such as *Sacchromyces cerevisiae* and *Caenorhabditis elegans* (in contrast to *Mus musculus* or *Rattus norvegicus*) were described as putative model organisms for human diseases [3].

This study was different from the approach of McGary et al. in the way that it depends on network structure of genetic pathways as well as Gene Ontology annotations. The work of McGary et al. was based on overlapping sets of orthologous genes, and a mathematical formulation based on these sets was used to find model organisms. McGary et al.'s work was based on molecular sequence information, without using network analysis to rank model organisms based on predicted subsystems (although McGary

Table 3
Number of interactions and proportions of predicted pathways that correctly matched reference pathways for a given species.

NCBI Taxon ID	Species name	Number of interactions			Proportions of predicted pathways matched to reference pathways	
		BioGRID	STRING	Total	Cancer	Infectious disease
3702	*Arabidopsis thaliana*[a]	13,828	0	13,828	0.0419	0.055
4896	*Schizosaccharomyces pombe*[b]	17,495	32,505	50,000	–	–
6239	*Caenorhabditis elegans*	6998	0	6998	0	0.004
7227	*Drosophila melanogaster*	40,153	9848	50,001	0.006	0.012
7955	*Danio rerio*	112	47,029	47,141	0.171	0.144
9031	*Gallus gallus*	180	39,337	39,517	0.067	0.063
9598	*Pan troglodytes*	0	36,756	36,756	0.158	0.153
9615	*Canis lupus familiaris*	5	33,398	33,403	0.014	0.014
9823	*Sus scrofa*[b]	1	12,831	12,832	–	–
9913	*Bos taurus*	33	49,967	50,000	0.614	0.192
10090	*Mus musculus*	4729	45,271	50,000	0.452	0.506
10116	*Rattus norvegicus*	851	49,163	50,014	0.391	0.289
511145	*Escherichia coli*[c]	4	49,996	50,000	0.236	0.037
559292	*Saccharomyces c. S288c*[c]	234,870	0	234,870	0.023	0.012

[a] Reference pathways from AraPath database.
[b] Species without a WikiPathways entry.
[c] *Escherichia coli* and *Saccharomyces c. S288c* were compared to BioCyc reference pathways. All predictions regarding other species were evaluated using WikiPathways.

et al. studied connectivity and modularity of the subsystems they discovered in cellular networks of candidate organisms, but that was a further analysis step of the results and was not a core part of their described method).

Based on the results shown in Tables 4 and 5, it was also observed that performance of *Mus musculus* and *Rattus norvegicus* models was greatly different in the case of some cancer diseases (e.g., Renal cell carcinoma and Melanoma) and infectious diseases (e.g., Pertussis and Epstein-Barr Virus). These results suggest that it may be worth exploring *Danio rerio* (for Renal cell carcinoma, Melanoma, or Pertussis) or *Saccharomyces cerevisiae* (for Epstein-Barr Virus) as better disease models for certain diseases. To further support this finding, recent studies have proposed *Danio rerio* as a potential model organism for cancer [28–30], infectious and immune diseases [31], and *in vivo* drug discovery [32]. Furthermore, some genes of *Saccharomyces cerevisiae* have shown similarity to Epstein-Barr virus DNA polymerase and be orthologous to human genes associated with Epstein-Barr virus [33,34]. However, it is important to note that the plausibility of alternative model organisms might also require the consideration of other features such as phenotypic properties of these specific diseases (e.g., do the organisms exhibit an observable disease state phenotype that is alterable?) as well as other practical considerations (e.g., availability of valid wild-types or appropriate inbred species).

4.2. Choice of data resources and annotation scheme

Combining micro-level, molecular function annotations of gene/protein nodes together with information about semantics inherited in a graph structure can be a powerful approach to derive new findings of relevance to biomedicine. Node annotations might not be restricted to molecular function annotations of GO. Genes/proteins in pathways and interaction networks with disease-specific annotation could be augmented from a variety of knowledge sources. For example, it may be possible to leverage biobanking and phenotypic information from Electronic Health Records (EHR) [35] and clinical data resources to annotate disease genes/proteins. Indeed, we are currently exploring the potential to do this in the future, with the goal to develop an EHR knowledge-enriched model to study disease genes/proteins in the context of real clinical scenarios.

While GO annotations can be found in gene ontology annotations (GOA) files of the Gene Ontology database, HPRD was chosen as a source of GO annotations because it is a manually curated resource and GO-compatible database. HPRD initially started with data from the Online Mendelian Inheritance in Man (OMIM) database [36] that focused on disease related genes [37]. This level of curation met the scope of this study to learn knowledge from disease-related genetic pathways.

This study only made use of GO molecular function terms. GO biological process terms are more diverse (and more specific) in characterizing genes/proteins than molecular function terms (there are nearly 2.5 times more biological process terms than molecular function terms). For the purposes of this study, molecular function terms were able to increase the model generalization (extracted patterns can be matched to GO-annotated interaction networks), thus not requiring additional biological process terms. Even though the GO biological process terms were not used in the model, the KEGG edge annotations (e.g., general process type such as PPrel and specific relation types such as activation, expression and inhibition) do capture semantics of the biological process that involved two genes/proteins.

Using a major gene/protein interaction database such as BioGRID, which provides a high number of unique interactions among other major databases [38], can be a limiting factor for predicting subsystems in many species due to the low number of interactions for some species in BioGRID database. The use of gene/protein interactions drawn from meta-databases such as STRING enhanced the ability to recover known subsystems by increasing the size of interaction networks. The number of interactions (per species) imported from BioGRID and STRING databases highlights the importance of aggregating evidence information about interactions from large number of sources. For instance, the interaction network of *Escherichia coli* had only four interactions in BioGRID database. About 50,000 interactions regarding *Escherichia coli* imported from STRING enabled the prediction of 23% of Wikipathways reference pathways of *Escherichia coli*. For *Danio rerio*, the interaction network had only 112 interactions imported from BioGRID. Importing 47,029 interactions from STRING allowed for 18% prediction accuracy for cancer diseases class and 12% prediction accuracy of infectious diseases class. The majority of interactions imported from STRING regarding *Escherichia coli* and *Danio rerio* was largely supported by evidence scores from predicted interactions (e.g., genome context and text mining).

The contribution of multiple methods for interaction prediction can be demonstrated by the case of *Danio rerio* and *Escherichia coli* interaction networks constructed using interactions imported from STRING. As shown in Figs. 6 and 7, about 55% of interaction network constructed for the *Danio rerio* and about 80% of interaction network constructed for the *Escherichia coli* were derived from

Table 4
Detailed performance analysis of 14 Species on cancer diseases fingerprints. Each entry represents proportions of correctly predicted pathways for each species using fingerprints of the indicated disease pathway. These proportions were calculated by normalized through dividing correctly predicted pathways of a given species by the number of all correctly predicted pathways for the same disease. Bold values indicate best model organism in covering fingerprints of a disease.

NCBI Taxon ID	Species name	Cancer Pathway (KEGG ID)											
		Colorectal cancer (hsa05210)	Renal cell carcinoma (hsa05211)	Pancreatic cancer (hsa05212)	Endometrial cancer (hsa05213)	Glioma (hsa05214)	Thyroid cancer (hsa05216)	Basal cell carcinoma (hsa05217)	Melanoma (hsa05218)	Bladder cancer (hsa05219)	Chronic myeloid leukemia (hsa05220)	Acute myeloid leukemia (hsa05221)	Non-small cell lung cancer (hsa05223)
3702	*Arabidopsis thaliana*	0.018	0.005	0.001	0.004	0	0.014	0	0	0	0	0	0
4896	*Schizosaccharomyces pombe*	0	0	0	0	0	0	0	0	0	0	0	0
6239	*Caenorhabditis elegans*	0	0	0	0	0	0	0	0	0	0	0	0
7227	*Drosophila melanogaster*	0.004	0.001	0	0.002	0	0.001	0	0	0	0	0	0
7955	*Danio rerio*	0.04	**0.459**	0.059	0.226	0.07	0.08	0.015	**0.449**	0.215	0.228	0.077	0.098
9031	*Gallus gallus*	0.001	0.004	0	0.031	0	0.012	0	0	0	0	0	0
9598	*Pan troglodytes*	0.03	0.049	0.007	0.084	0.156	0.064	0	0.094	0	0.07	0.092	0.096
9615	*Canis lupus familiaris*	0.005	0.004	0	0.019	0	0.011	0	0	0	0.008	0.01	0.014
9823	*Sus scrofa*	0	0	0	0	0	0	0	0	0	0	0	0
9913	*Bos taurus*	0.002	0	0	0.008	0.001	0	0	0	0	0	0	0
10090	*Mus musculus*	**0.669**	0.25	0.395	0.28	0.344	**0.416**	**0.859**	0.384	**0.439**	0.296	0.314	0.351
10116	*Rattus norvegicus*	0.184	0.229	**0.536**	**0.348**	**0.428**	0.402	0.122	0.072	0.346	**0.398**	**0.508**	**0.441**
511145	*Escherichia coli*	0.022	0	0	0	0	0	0	0	0	0	0	0
559292	*Saccharomyces c. S288c*	0.025	0	0.002	0	0	0	0.005	0	0	0	0	0

Table 5
Detailed performance analysis of 14 Species on infectious diseases fingerprints. Each entry represents proportions of correctly predicted pathways for each species using fingerprints of the indicated disease pathway. These proportions were calculated by normalized through dividing correctly predicted pathways of a given species by the number of all correctly predicted pathways for the same disease. Bold values indicate best model organism in covering fingerprints of a disease.

NCBI Taxon ID	Species name	Infectious Disease Pathway (KEGG ID)															
		Bacterial invasion epithelium (hsa05100)	Helico-bacter pylori infection (hsa05120)	Esche-richia coli infection (hsa05130)	Shigellosis (hsa05131)	Pertussis (hsa05133)	Legionel-losis (hsa05134)	Leishman-iasis (hsa05140)	Chagas disease (hsa05142)	Toxoplas-mosis (hsa05145)	Tubercu-losis (hsa05152)	Hepatitis C (hsa05160)	Measles (hsa05162)	Influenza A (hsa05164)	HTLV-I (hsa05166)	Herpes simplex (hsa05168)	Epstein-Barr virus (hsa05169)
3702	*Arabidopsis thaliana*	0.005	0	0	0	0	0	0.006	0.004	0.001	0.002	0.022	0.003	0.021	0.002	0.009	0
4896	*Schizosaccharo-myces pombe*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6239	*Caenorhabditis elegans*	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7227	*Drosophila melanogaster*	0.008	0	0	0	0	0	0.002	0.001	0	0	0.002	0.001	0	0	0.001	0
7955	*Danio rerio*	0.019	0	0	0.068	**0.348**	**1**	0.248	0.136	0.012	0.067	0.161	0.103	0.046	0.065	0.05	0
9031	*Gallus gallus*	0.005	0	0	0.002	0.001	0	0.013	0.008	0.005	0.006	0.024	0.002	0.001	0.002	0	0
9598	*Pan troglodytes*	0.019	0	0	0.058	0.158	0	0.083	0.038	0.072	0.055	0.081	0.046	0.033	0.064	0.007	0
9615	*Canis lupus familiaris*	0.003	0	0	0.013	0	0	0.006	0.003	0.002	0.004	0.009	0.001	0.001	0.004	0.003	0
9823	*Sus scrofa*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9913	*Bos taurus*	0.003	0	0	0	0	0	0	0.001	0.007	0.001	0.001	0	0	0	0.001	0
10090	*Mus musculus*	**0.693**	**1**	**1**	**0.638**	0.211	0	**0.346**	**0.617**	**0.624**	**0.66**	**0.39**	**0.596**	**0.65**	**0.588**	**0.793**	0.171
10116	*Rattus norvegicus*	0.227	0	0	0.221	0.282	0	0.296	0.189	0.265	0.202	0.31	0.247	0.245	0.274	0.125	0
511145	*Escherichia coli*	0.004	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
559292	*Saccharomyces c. S288c*	0.014	0	0	0	0	0	0	0.004	0.011	0.002	0	0.001	0.005	0	0.01	**0.829**

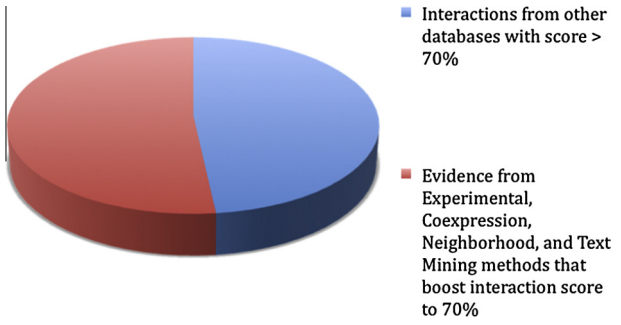


Fig. 6. Contribution of methods used to predict interactions for *Danio rerio*.

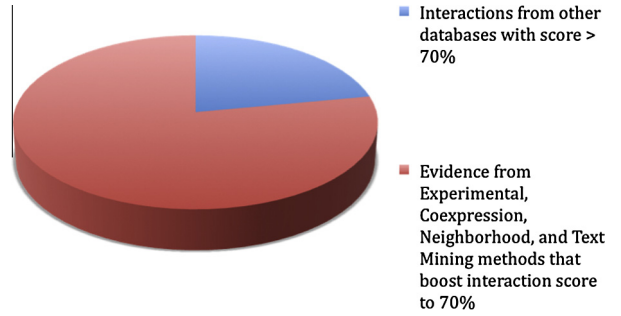


Fig. 7. Contribution of methods used to predict interactions for *Escherichia coli*.

evidence from experimental, gene expression, text mining, and gene neighborhood methods that collectively increased the overall evidence score above 70%. As has been done by others (e.g., Ferrer et al. [8] used threshold of >0.5 for an adjusted rand index for determining the correctness of a pathway), a threshold of 70% was mainly chosen to imply that more than two thirds of the genes/proteins in a pathway are found. However, if the *Danio rerio* and *Escherichia coli* networks were constructed only from data imported from major databases, the networks would respectively be 45% and 20% of their potential size. Table 6 shows statistics about the STRING interactions used in the construction of the *Danio rerio* network. While 98% of *Danio rerio* network links had non-zero scores for partial evidence derived from other databases, 55% these partial evidence scores would not pass the 70% threshold and hence the *Danio rerio* networks would be 45% of its size. Partial evidence from experimental, gene expression, text mining, and gene neighborhood methods thus boosted the size of *Danio rerio* network. The results shown in Table 4 also suggest that, for some species, very few known interactions (118 as in the *Danio rerio* dataset) were available in BioGRID database. Including interactions from STRING (mostly predicted interactions) allowed for a wider coverage of the interaction network. The overall impact of including multiple sources resulted in an improvement of overall prediction accuracy for 18% of subsystems discovered by cancer fingerprints and 12% for infectious disease fingerprints.

Table 6
Interactions of *Danio rerio* interaction network with detailed sources of evidence.

Evidence Method/ Source	Number of STRING links with non-zero score	Proportion of network links with non-zero score (%)
Neighborhood	4891	9.7
Fusion	299	0.6
Cooccurence	2219	4.4
Coexpression	22,319	44
Experimental	11,547	23
Other databases	49,103	98
Text Mining	15,149	30

4.3. Summary of study contributions

There are four major contributions of the methodology developed in this study for evaluating potential model organisms. First, it was shown that a model-based method could be used to search and extract functional structural patterns (disease fingerprints) in disease pathway graphs. Second, a subgraph pattern matching algorithm, supported by a simple and memory-efficient indexing method was shown to be useful for identifying subsystems in interaction networks using disease fingerprints. Third, this work leveraged rich knowledge sources (KEGG pathways, BioGRID and STRING interactions databases that could be annotated with GO) together with computational mining methods to infer potentially new knowledge (e.g., novel subsystems of disease). The fourth, and perhaps most significant, way that the methodology presented here is different from previous studies is that the assessment of disease model potential was achieved at both the unit level (by considering molecular function) and system level (by considering graph structure patterns in pathways). Thus, this approach is different from related studies that used gene orthologue sets as the basis to assess how an organism was suitable as a model (e.g., most recently by McGary et al. [3]). The method used in this study complements these types of approaches in two major ways: (1) the way gene molecular function is used to represent similarity of genes in different organisms and (2) pathways are predicted using system-level graph-based methods.

4.4. Study limitations

The methods presented here have a number of limitations related to decisions about the computational methods, the data resources, and the assumptions made in this study. The EM algorithm that was used for parameter estimation (see Appendix A) is known for not guaranteeing optimum solutions. Graphs in the KEGG pathway datasets were assumed to be independent and identically distributed data. While it is hard to confirm that a given pair of pathways sharing a set of genes/proteins is totally independent, assuming independence of graphs items was for the purpose of statistical analysis and to make the computation of the model more tractable. There are a number of alternative resources that might have been used, including Reactome pathways and molecular networks [39], species-specific databases such as Rat Genome Database (RGD) [27] and WormBase [40].

Some limitations are inherent in the datasets chosen for this study and could have had an impact on the results produced. For instance, significant proportions of the interactions in STRING database are predicted interactions and thus there is always a possibility of errors about predicting two genes/proteins being genetically or physically interacting. There might be gene set overlap between pathway data from Wikipathways, BioCyc, and AraPath databases with the KEGG human pathways that were used as design dataset. However, this did not have a significant effect on quality of evaluation procedure for two reasons. First, the method described in this study did not use any sequence similarity or homology-based technique to predict pathways similar to those of humans in other species. Second, the methodology used in this study relied on the molecular function of genes, not the genes themselves and therefore, genes in predicted pathways did not necessarily have to be sequence-based homologues of human genes.

Evaluating candidate model organisms at the molecular level is only one facet for determining the viability of a possible model organism. Other factors, such as cost and controllability in a lab environment, also need to be considered. This study aimed to utilize already available resources about potential model organism for systematic evaluation, without particular consideration of cost or controllability. Nonetheless, the use of the approach described in

this study may be one factor that can be combined with cost and controllability factors to help guide future research on human diseases.

5. Conclusion

This study proposed a method for the evaluation of species as models to study human diseases. Disease-related genetic pathways were functionally and structurally analyzed to uncover characteristic subgraph patterns. These patterns were then matched to molecular interaction networks for 14 potential model organisms. The adequacy of a given species as a potential disease model was hypothesized to be related to the degree to which interaction networks cover disease patterns. The finding that proportions of correctly predicted subsystems in *Danio rerio* (Zebrafish) and *Saccharomyces cerevisiae* (Baker's yeast) interaction networks were higher than those of two common model organisms *Mus musculus* (Mouse) and *Rattus norvegicus* (Rat) suggests there might be unobvious molecular networks in alternative model organisms that might be relevant to study disease-related processes. The findings of this study suggest that a network, system-level approach can be an effective means to find such unobvious networks. The promising results of this study suggest that the disease fingerprint approach may be used to analyze pathways across multiple species and may thus be used to identify model organisms for the study of human disease related processes.

Appendix A. Model and algorithm for subgraph pattern analysis of genetic pathways

A.1. Preliminaries and Notations

A graph G consists of a set of vertices V and a set of edges E in which each edge $e \in E$, denoted by $e(u,v)$, links two vertices u, $v \in V$. A subgraph consists of a set of nodes $V' \subseteq V$ together with a set of edges $E' \subseteq E$ that links its nodes. In this study, genetic pathways were modeled as a set of labeled directed graphs.

Definition A1 (*Labeled Graph*). A labeled graph $G(V,E,L_V,L_E,\sum_V,\sum_E)$ has a node labeling function L_V: $V \to \sum_V$ that assigns labels from a node alphabet set $\sum_V$ to nodes and an edge labeling function L_E: $E \to \sum_E$ that assigns labels from an edge alphabet set $\sum_E$ to edges. A labeled subgraph g consists of a subset of nodes of G and edges that link them. Labels of nodes and edges of a subgraph g are the same as its super graph G.

The node alphabet set $\sum_V$ contained GO terms. The edge alphabet set $\sum_E$ contained relation types in KEGG disease pathway dataset. The basic definition of a labeled graph was extended in two ways. First, a NULL label ε was assigned to gene nodes with no GO terms associated. Second, an entity could be mapped to more than one label. Also, edges in a given pathway could be labeled with more than one relation type.

Disease fingerprints are subgraphs of GO annotated disease pathways graphs that were are assumed to represent functional sub-processes that could be characteristics of a disease class such as immune, infectious, or neurodegenerative disease. Disease fingerprints are therefore functional structural patterns in GO annotated graphs. To quantify the degree to which a fingerprint was related to each disease class, a first step was to use a utility function to highlight a set of subgraphs (fingerprints) in a given pathway graph. This utility function was termed a "partitioning function."

Definition A2 (*Partitioning Function*). Let $E(\cdot)$ denote edge set of a graph G. A partitioning function π: $E(G) \to Z$ assigns an integer to every edge of G such that edges with the same integer form a subgraph. The set of subgraphs $H\pi$ highlighted by a specific partitioning function π is defined as $H\pi = \{g_i | \forall e \in E(g_i),\ E(g_i) \subseteq E(G),\ \pi(e) = i\}$.

Fig. 2 illustrates the concept of partitioning. From the above definition, it follows that every edge must be covered by only one subgraph (i.e., subgraphs of a given partitioning are edge-disjoint). There is a big space of partitioning functions for each graph in the dataset and this space is not known *a priori*. Searching for good partitioning functions was thus one of the objectives of this study. Preventing subgraph overlapping has a useful impact on speed and memory during search for partitioning for each graph. For instance, the probability estimation algorithm (presented in the next section) does not have to minimize overlapping of subgraphs while searching for partitionings. To accommodate side effects of this restriction, the process of identifying disease fingerprints takes into account information from many hypothesized partitioning functions for every graph in the dataset. In this study, partitionings were represented by integer arrays where indices represent edge identifiers and values represent subgraphs to which an edge belongs. This compact representation allowed for easy extension of partitionings by modifying edge-to-subgraph assignments of an existing partitioning in order to generate new ones. This array representation was also helpful in detecting similarity between partitionings (which was useful in minimizing memory requirements of the application by keeping only one copy of a partitioning among several equi-probable partitionings).

A.2. Mathematical model

Graphs in the design dataset were assumed to be independent and identically distributed (*iid*) data observed from an unknown probability distribution $P(G)$. The *iid* data assumption was made for the purpose of facilitating statistical inference and to make decision about properties (e.g., class label) of a graph instance independent of other graph instances in the dataset. For each pair of graph G and disease class C, a probability value was used to quantify the relation between a graph and its class label. Let the probability value $P(G|C)$ quantify the characteristics of class C that is observed in graph G. Modeling this probability value directly can be hard, mainly because: (1) it is a computationally non-trivial task to determine if two graph instances are equal using the graph isomorphism test [19,41]; and (2) due to the data sparseness problem (it is usually hard to find more than one isomorphic instance of the same graph in a given dataset). An indirect way to model $P(G|C)$ was used to provide the model with access to GO functional annotations as well as hidden structural patterns (collectively referred to as 'fingerprints') in a given graph. Using the utility of partitioning function, a more useful probability value $P(G,\pi|C)$ would involve a graph instance G, a class label C, and a graph partitioning π that divides G into a set of fingerprints. $P(G,\pi|C)$ quantifies the probability of observing structural patterns of class C in graph instance G. There are many possible partitionings for the same graph instance, and to take into account all possible structural patterns represented in these partitionings, the probability value $P(G|C)$ can be expressed as

$$P(G|C) = \sum_{\pi} P(G,\pi|C) \tag{A1}$$

Let $H\pi$ be the set of subgraphs according to a partitioning function π of graph G:

$$H\pi = \{g_i | \forall e \in E(g_i),\ E(g_i) \subseteq E(G),\ \pi(e) = i\}$$

Assuming that subgraphs resulting from a partitioning function are conditionally independent, $P(G,\pi|C)$ can be written as

$$P(G,\pi|C) \prod_{g \in H\pi} P(g|C) \tag{A2}$$

The probability value $P(g|C)$ represents the likelihood that subgraph g is a characteristic structural pattern of class C. Here, it should be pointed out that the conditional independence assumption made here is mathematically plausible considering that: (1) subgraphs in one partitioning do not overlap (i.e., do not share common edges, according to definition of π); and (2) this assumption is made for subgraphs within the same partitioning (i.e., it is local to a specific partitioning, not for all combinations of subgraphs.) For the purpose of probability estimation, counting the number of instances of a subgraph in all partitionings of graph dataset is impractical, since it re-introduces the problem of subgraph isomorphism [19]. In this study, GO-annotated maximal paths inside the subgraphs were used to approximate representation of subgraphs. Each maximal path represented a sequence of GO annotations of nodes that lay in that maximal path. In case a node has more than one GO annotation, multiple maximal paths are generated so that each maximal path has only one GO annotation per node. Then, $P(g|C)$ was calculated approximately as

$$P(g|C) \approx \prod_{a \in g} P(a|C) \tag{A3}$$

where a denotes a GO-annotated maximal path that connect a subset of nodes inside subgraph g. Using Eq. (A3), the likelihood of a partitioning and a graph instance given a disease class label can be written as

$$P(G, \pi|C) = \prod_{g \in H \pi} \prod_{a \in g} P(a|C) \tag{A4}$$

and, finally, $P(G|C)$ can be expressed as

$$P(G|C) = \sum_{\pi} \prod_{g \in H \pi} \prod_{a \in g} P(a|C) \tag{A5}$$

Thus, Eqs. (A2)–(A5) casts the problem of searching for disease fingerprints as estimating a conditional distribution of GO annotated maximal paths given disease classes, while maintaining a set of best partitionings for each graph instance highlighting disease fingerprints.

A.3. Probability estimation and searching for best partitionings

For a given pathway design dataset, two data entities need to be generated: (1) the best scoring partitioning set (that contains disease fingerprints within each pathway); and (2) the conditional probability table $P(a|C)$. The generation of each of these two entities requires the existence of the other, but neither of them exists with the graph data at the beginning of probability estimation process. Therefore, both entities must be generated initially at the same time, albeit with low likelihood, and probability estimate of $P(a|C)$ and partitionings likelihood values can be improved iteratively. Here, the estimation of model parameters θ is performed following a maximum likelihood approach using the Expectation–Maximization (EM) [20,42] algorithm. Model parameters consisted of the probability distribution of maximal paths given class labels:

$$\theta = \{P(a|C)\} \tag{A6}$$

There can be a large space of possible values of the parameters θ and the search for best parameter values can be based on maximizing the likelihood on the graph dataset:

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \left\{ \prod_{n=1}^{N} [P_{\theta}(G_n|C_n)] \right\} \\ &= \arg\max_{\theta} \left\{ \prod_{n=1}^{N} \left[\sum_{\pi} P_{\theta}(G_n, \pi|C_n) \right] \right\}, \end{aligned} \tag{A7}$$

where N is the number of graphs in the dataset and $P_{\theta}(G_n, \pi|C_n)$ is computed by Eq. (A4) and the probability distribution P_{θ} $(a|C)$. $P_{\theta}(G_n, \pi\ |C_n)$ represents the probability of a partitioning of a graph given a class label using a given set of values of parameters θ. The EM algorithm aims at maximizing the likelihood function in Eq. (A7) while identifying best graph partitionings that highlight key patterns. Because it was computationally expensive to consider all possible partitionings for graphs in the probability estimation algorithm, a priority queue of a limited number of highly probable partitionings was maintained. In each iteration, searching for new partitionings extends the set of best partitionings of each graph. These partitionings are evaluated using Eq. (A4) and the parameter values θ obtained in the previous iteration. An initial set of random partitionings is generated for each graph in the dataset. Annotated maximal paths were extracted from each subgraph of a given partitioning and the parameters θ are initialized with uniform probability values. The EM algorithm consisted of repeated iterations of E-Step and M-Step. In the E-Step of the algorithm, and for each graph, maximal path parameter counts are collected from within partitionings. The count of a parameter in one graph is calculated using:

$$c(a|C; G) = \sum_{\pi} P(\pi|G, C) N(a, G) \sum_{j} \delta(a, a_j) \delta(C, C_j) \tag{A8}$$

Here, $N(a,G)$ is the number of times a maximal path a appeared in G (in different subgraphs of G), and δ is the Kronecker's delta function. The probability value $P(\pi|G,C)$ is the normalized partitioning probability conditioned on a graph and a class and is given by:

$$P(\pi|G, C) = \frac{P(\pi, G|C)}{\sum_{\pi'} P(\pi', G|C)} \tag{A9}$$

where $P(\pi',G|C)$ is given by Eq. (A4). The summation in Eq. (A9) is over the set of best partitionings that is generated for graph G. Since this set is limited in size, Eq. (A9) is only an approximation of partitioning quality. Multiplying the path-class counts $\delta(a,a_j)\delta(C,C_j)$ by partitioning probability $P(\pi|G,C)$ in Eq. (A8) aimed at weighing each parameter count according to partitionings quality represented by $P(\pi|G,C)$. In the M-step, the maximal path parameters are computed by normalizing the counts:

$$P(a|C) = \frac{\sum_{n} c(a|C; G_n)}{\sum_{n,a} c(a|C; G_n)} \tag{A10}$$

For each iteration of the model training algorithm, a search for best partitionings is performed using the best parameters θ estimated so far. Generating new partitionings from existing partitionings can be achieved moving edges from one subgraph to another subgraph. This way some subgraph patterns can grow while others can diminish. Fig. 3 illustrates the process of generating a new partitioning from an existing one. Both existing and newly generated partitionings were evaluated using Eq. (A4) based on the most recent parameter values θ. In this study, the parameter estimation algorithm was run four iterations.

In summary, the model training procedure aimed to estimate the probability distribution $P(a|C)$. As a by-product of this procedure, a set of best partitionings of each graph highlighted the key subgraph patterns in the dataset. The pattern analysis model described above was used to find best partitionings in disease pathways with nodes annotated with molecular functions. Key patterns were extracted from best partitionings of pathways to be matched to sub-networks in gene/protein interaction network of a species.

Appendix B. Matching query subgraphs to an interaction network

Algorithm 1. Matching Subgraph Patterns to Gene/Protein Interaction Network

Input:

Network index: index, network adjacency matrix, subgraph pattern: *g*, *V(g)* vertex set of g

Process:

Step 1: Initialization

1: for each node v in $V(g)$
2: initialCandidateMatchingSet(v) = {}
3: for each neighbor node u of v
4: *mSet* = *{}*
5: *let* $k \leftarrow (L_v(v), L_v(u))$
6: vals = index.get(k)
7: for each $x \in$ vals
8: mSet.insert(x)
9: initialCandidateMatchingSet(v).insert(mSet)

Step 2: Applying topological constraints

10: for each node v in $V(g)$
11: let candidateMatchingSet(v) be intersection of all sets in initialCandidateMatchingSet(v)
12: for each node v in $V(g)$
13: let S = candidateMatchingSet(v)
14: remove every item $i \in S$ if i is not linked to any item of candidate sets of neighbors of node v
15: return *{}* if S is empty
16: repeat 15–17 until no item can be removed from candidate sets

Step 3: Generate subnetworks by finding edges between nodes in final candidateMatchingSet

17: *matchedSubNetworks* = Array(|*V*(*g*)|)
18: Let S be the array of all nodes in $V(g)$
19: *matchedSubNetworks* (*S[1]*) = {}
19: for each network node identifier u in candidateMatchingSet of node *S[1]*
20: *matchedSubNetworks* (*S[1]*).append({*u*})
21: for $i = 2$ to |*S*| // |S| denotes size of set S
22: *partialnetworks* = *matchednetworks*($S[i-1]$)
23: for each partial network h in *partialnetworks*
24: for each network node identifier u in *candidateMatchingSet* of node *S[i]*
25: if $\exists$ *node* $w \in h$ such that u is linked to w in the interaction network
25: *matchednetworks*($S[i]$).append($\{u\} \cup h$)

Output: matchedSubNetworks[|*S*|] // *output last element in the array matchedSubNetworks*

References

[1] Aitman TJ, Boone C, Churchill GA, Hengartner MO, Mackay TFC, Stemple DL. The future of model organisms in human disease research. Nat Rev Genet 2011;12:575–82.
[2] Thomas MA, Yang L, Carter BJ, Klaper RD. Gene set enrichment analysis of microarray data from *Pimephales promelas* (Rafinesque), a non-mammalian model organism. BMC Genomics 2011;12:66.
[3] McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc Natl Acad Sci 2010;107:6544.
[4] Bedell MA, Largaespada DA, Jenkins NA, Copeland NG. Mouse models of human disease. Part II: recent progress and future directions. Genes Dev 1997;11:11.
[5] Koteja P, Garland T, Sax JK, Swallow JG, Carter PA. Behaviour of house mice artificially selected for high levels of voluntary wheel running. Anim Behav 1999;58:1307–18.
[6] Haldar M, Hancock JD, Coffin CM, Lessnick SL, Capecchi MR. A conditional mouse model of synovial sarcoma: insights into a myogenic origin. Cancer Cell 2007;11:375–88.
[7] Haldar M, Randall RL, Capecchi MR. Synovial sarcoma: from genetics to genetic-based animal modeling. Clin Orthop Relat Res 2008;466:2156–67.
[8] Ferrer L, Shearer AG, Karp PD. Discovering novel subsystems using comparative genomics. Bioinformatics 2011;27:2478–85.
[9] Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA 2005;102:13544–9.
[10] Bebek G, Yang J. PathFinder: mining signal transduction pathway segments from protein–protein interaction networks. BMC Bioinformatics 2007;8:335.
[11] Cakmak A, Ozsoyoglu G. Mining biological networks for unknown pathways. Bioinformatics 2007;23:2775.
[12] Senf A, Chen X-W. Identification of genes involved in the same pathways using a Hidden Markov model-based approach. Bioinformatics 2009;25:2945–54.
[13] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 2010;38:D355.
[14] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–9.
[15] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25.
[16] Nabhan AR, Sarkar IN. Mining disease fingerprints from within genetic pathways. In: Annual AMIA symbosium. Chicago, IL; 2012.
[17] Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database – 2009 update. Nucleic Acids Res 2009;37:D767–72.
[18] Yan X, Han J. gSpan: graph-based substructure pattern mining. IEEE; 2002. p. 721–4.
[19] Read RC, Corneil DG. The graph isomorphism disease. J Graph Theory 1977;1:339–63.
[20] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol) 1977:1–38.
[21] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011;39:D561–8.
[22] Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 2008;36:D623–31.
[23] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28:235–42.
[24] Carbon S, Ireland A, Mungall CJ, Shu SQ, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. Bioinformatics 2009;25:288–9.
[25] Pico AR, Kelder T, Van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol 2008;6:e184.
[26] Lai L, Liberzon A, Hennessey J, Jiang G, Qi J, Mesirov JP, et al. AraPath: a knowledgebase for pathway analysis in Arabidopsis. Bioinformatics 2012;28:2291–2.
[27] Dwinell MR, Worthey EA, Shimoyama M, Bakir-Gungor B, DePons J, Laulederkind S, et al. The Rat Genome Database 2009: variation, ontologies and pathways. Nucleic Acids Res 2009;37:D744–9.
[28] Stoletov K, Klemke R. Catch of the day: zebrafish as a human cancer model. Oncogene 2008;27:4509–20.
[29] Stern HM, Zon LI. Cancer genetics and drug discovery in the zebrafish. Nat Rev Cancer 2003;3:533–9.
[30] Feitsma H, Cuppen E. Zebrafish as a cancer model. Mol Cancer Res 2008;6:685–94.
[31] Sullivan C, Kim CH. Zebrafish as a model for infectious disease and immune function. Fish Shellfish Immunol 2008;25:341–50.
[32] Zon LI, Peterson RT. In vivo drug discovery in the zebrafish. Nat Rev Drug Discov 2005;4:35–44.
[33] Morrison A, Christensen R, Alley J, Beck A, Bernstine E, Lemontt J, et al. REV3, a *Saccharomyces cerevisiae* gene whose function is required for induced mutagenesis, is predicted to encode a nonessential DNA polymerase. J Bacteriol 1989;171:5659–67.
[34] Dheekollu J, Lieberman PM. The replisome pausing factor timeless is required for episomal maintenance of latent Epstein-Barr virus. J Virolgy 2011;85:5853–63.
[35] Jensen PB, Jensen LJ, Brunak SR. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13:395–405.
[36] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005;33:D514–7.
[37] Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 2003;13:2363–71.
[38] Lehne B, Schlitt T. Protein–protein interaction databases: keeping up with growing interactomes. Hum Genomics 2009;3:291–7.
[39] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 2005;33:D428–32.
[40] Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, et al. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res 2010;38:D463–7.
[41] Shang H, Zhang Y, Lin X, Yu JX. Taming verification hardness: an efficient algorithm for testing subgraph isomorphism. Proc VLDB Endowment 2008;1:364–75.
[42] Moon TK. The expectation-maximization algorithm. Signal Process Mag, IEEE 1996;13:47–60.