Molecular BioSystems

RSC Publishing

PAPER

View Article Online
View Journal

A disease-drug-phenotype matrix inferred by walking on a functional domain network†

Cite this: DOI: 10.1039/c3mb25495j

Hai Fang*‡ and Julian Gough*‡

Received 1st November 2012,
Accepted 26th February 2013

DOI: 10.1039/c3mb25495j

www.rsc.org/molecularbiosystems

Protein domains are classified as units of structure, evolution and function, and thus form the molecular backbone of biosphere. Although functional networks at the protein level have been reported to be of value in predicting diseases (phenotypes or drugs), they have not previously been applied at the sub-protein resolution (protein domain in this case). We herein introduce a domain network with a functional perspective. This network has nodes consisting of protein domains (at the superfamily/evolutionary level), with edges weighted by the semantic similarity according to domain-centric Gene Ontology (dcGO) annotations, which henceforth we call ''dcGOnet''. By globally exploring this network *via* a random walk, we demonstrate its predictive value on disease, drug, or phenotype-related ontologies. On cross-validation recovering ontology labels for domains, we achieve an overall area under the ROC curve of 89.0% for drugs, 87.3% for diseases, 87.6% for human phenotypes and 88.2% for mouse phenotypes. We show that the performance using global information from this network is significantly better than using local information, and also illustrate that the better performance is not sensitive to network size, or the choice of algorithm parameters, and is universal to different ontologies. Based on the dcGOnet and its global properties, we further develop an approach to build a disease-drug-phenotype matrix. The predicted interconnections are statistically supported using a novel randomization procedure, and are also empirically supported by inspection for biological relevance. Most of the high-ranking predictions recover connections that are well known, but others uncover connections that have only suggestive or obscure support in the literature; we show that these are missed by simpler methods, in particular for drug-disease connections. The value of this work is threefold: we describe a general methodology and make the software available, we provide the functional domain network itself, and the ranked drug-disease-phenotype matrix provides rich targets for investigation. All three can be found at http://supfam.org/SUPERFAMILY/dcGO/dcGOnet.html.

Introduction

Network biology[1] provides a conceptual framework for biomedical research at the level of molecular interactions. Protein interaction networks are the mainstay of this field.[2–4] Since protein–protein physical interaction data tend to be error-prone and far from complete,[5,6] the use of functional aspects of interactions is becoming more and more popular for a variety of applications, such as phenotype prediction,[7] genetic modifier loci discovery,[8] and disease gene prioritization as well as drug target identification.[9] The common principle is that: functionally related genes (or proteins) tend to be associated with the same or similar diseases (drugs or phenotypes). Translating this into the language of networks: genes (or proteins) causing the same or similar diseases (drugs or phenotypes) are more likely to lie closer in a functional protein network, which thus can be used for predicting diseases (drugs or phenotypes).

Protein domains, classified as units of structure, evolution and function by the Structural Classification of Proteins (SCOP)[10] database, represent direct manifestations of molecular biosphere. Inspired by the multifaceted utilities of functional networks of whole proteins, we hypothesize that functional networks at a sub-protein domain resolution may also be of great value and utility. One of the standout advantages of using domains rather than proteins is that: having domains as nodes can bypass many of the problems inherent in cross-knowledge

Department of Computer Science, University of Bristol, The Merchant Venturers Building, Bristol BS8 1UB, UK. E-mail: hfang@cs.bris.ac.uk, gough@cs.bris.ac.uk

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb25495j

‡ Author contributions: HF conceived, designed and implemented the experiments, analysed the results, and drafted the manuscript. JG conceived and guided the experiments, interpreted the results, and finalised the manuscript. All authors read and approved the final manuscript.

and cross-species comparisons when using protein networks directly. Another advantage of using domains is that in some cases treating the domain as the functional unit reveals deeper biological insights than simply collapsing all component functions to the whole protein and ignoring common components between proteins. Edges in domain networks are usually computationally inferred either from the protein interactions[11] or *via* domain co-occurrence in genomes.[12,13] The former assumes a structural basis for the protein interactions, while the latter can be used for studies on the genome organization and evolution. Both of them, however, are largely context-dependent.

To our knowledge, there is a lack of domain networks that are constructed with a functional perspective, and that are not reliant on purely context-specific information. Recent progress makes it feasible to fill this gap. The domain-centric Gene Ontology (dcGO) database[14] provides associations between GO terms and protein domains, such that each domain has a list of associated terms. In a similar way to measuring protein similarity *via* GO,[15] the degree of functional relatedness between domains can be assessed by measuring the semantic similarity of their respective GO annotation profiles. In dcGO, biomedical ontologies other than GO are also provided, *e.g.*, those created to describe diseases, drugs, and phenotypes in human and mouse. In this study we illustrate the power of a dcGO-derived functional domain network (hereinafter referred to as 'dcGOnet') by generating a disease-drug-phenotype matrix. To do this, we implemented a parallel version of a random walk with restart algorithm (RWR; originally proposed for image segmentation[16]) on this domain network. A RWR over the network enables a variable aggregation of relatedness information. Based on pre-computed pair-wise affinity scores for domains, we conducted a leave-one-out cross-validation test to evaluate the ability to recover drug, disease or phenotype ontology labels using the dcGOnet. We also devised a new approach that we used to build a disease-drug-phenotype connectivity matrix. The interconnections within the matrix are statistically significant as assessed by a degree-approximating randomization procedure.

Methods

Domain-centric annotations of functions, diseases, phenotypes and drugs

The latest release of the dcGO database[14] contains protein domain annotations with GO[17] and many other commonly used biomedical ontologies[18] including diseases, phenotypes, drugs and so forth. The focus in dcGO is on domains taken from the SCOP database,[10] although other domain databases are also annotated. In this study we use SCOP domains classified at the superfamily level (defined as grouping together domains for which there is structure, sequence and function evidence for a common ancestor). The domain-centric annotations are statistically inferred from proteins with experimental evidence,[19] and intuitively they can be understood as the modes-of-action underlying the protein. The dcGO algorithm takes as inputs lists of proteins annotated by an ontology term and associates common domains within them to this term along with probabilistic scores. A GO term can annotate many domains, and a domain can be annotated by many GO terms. These many-to-many relationships enable a more complete and less subjective description of functional similarity between domains (see the 'Constructing the functional domain network' Section). Similar to GO, the other ontologies including Disease Ontology (DO),[20] Human Phenotype (HP),[21] Mouse Phenotype (MP)[22] and DrugBank ATC codes (DB)[23] are organised hierarchically; terms higher up are more general and can annotate much more domains. These GO-independent ontologies are used for testing the ability of the functional domain network in recovering the ontology labels, avoiding the circularity of a GO-only assessment on the same terms from which the network was derived (see the 'Performance evaluation using leave one-out cross-validation' Section).

Constructing the functional domain network

We constructed a functional domain network (dcGOnet) by measuring the semantic similarity between domains according to their GO annotations. Specifically, the information content (IC) for a GO term is first quantified as the negative logarithm of the frequency of domains being annotated to this term. The IC tells how informative/specific a term is, with a higher value indicating a more informative/specific term (with respect to domain annotations). Then, the semantic similarity between two GO terms, t_1 and t_2, is quantified as the highest IC among all common ancestors that two terms share, *i.e.*, the Most Informative Common Ancestor, MICA(t_1, t_2). Since a domain can be annotated by multiple GO terms, the similarity measures between two domains should take into account all possible pair-wise term similarities. We adopted the ''best-match average'' method[15] to calculate the semantic similarity SIM(d_1, d_2) between two domains, d_1 and d_2:

$$\mathrm{SIM}(d_1,d_2) = \frac{1}{n_1+n_2}\left(\sum_{t_1\in T_1}\underset{t_2\in T_2}{\mathrm{MAX}}(\mathrm{MICA}(t_1,t_2)) + \sum_{t_2\in T_2}\underset{t_1\in T_1}{\mathrm{MAX}}(\mathrm{MICA}(t_1,t_2))\right), \quad (1)$$

where T_1 and T_2 stand for a set of multiple terms that annotate d_1 and d_2, respectively, n_1 for the size of T_1, and n_2 for the size of T_2. We calculate functional similarity for every domain pair, and construct a network with domains as nodes and functional similarity as the edge weights. The final network was obtained by selecting the top 50% of edges with the highest weights, resulting in dcGOnet containing 289 528 edges between 1075 nodes. It should be noted that this final network used for the subsequent analysis is still weighted, *i.e.*, it is weighted network but with the lowest 50% of edges removed. Notably, even the complete network is not fully connected, consisting of 1135 nodes but only 579 056 edges (there are 643 545 possible edges); this is because both domains of a pair must be annotatable by GO terms for there to be a non-zero edge weight.

Implementing the random walk with restart on the functional domain network

Conceptually similar to Google's PageRank algorithm,[24] the random walk with restart (RWR) algorithm iteratively explores the global structure of the network to estimate the proximity (affinity score) between two nodes. Starting at a node s, the walker faces two choices at each step: either moving to a randomly chosen neighbor, or jumping back to the starting node s. The algorithm is formulated as:

$$\vec{p}_s(t+1) = (1-r) \times \tilde{W} \times \vec{p}_s(t) + r \times \vec{p}_s(0) \qquad (2)$$

where $\tilde{W}$ is the normalized Laplacian adjacency matrix W associated with the weighted network, r is a fixed parameter for the restarting probability (with $1-r$ for the probability of moving to a neighbor), $\vec{p}_s(0)$ is the starting probability vector with 1 for the starting node s and 0 for the rest, and $\vec{p}_s(t)$ denotes the probability vector that the walker visits network nodes at step t. After iteratively reaching stability (as measured by $\|\vec{p}_s(t+1) - \vec{p}_s(t)\|_1 < 10^{-6}$), the steady probability vector $\vec{p}_s(\infty)$ contains the affinity score of all nodes in the network to the starting node s. This steady probability vector can be viewed as the "influential impact" over the network imposed by the starting node.

The normalized Laplacian of the network/graph[25] ensures that the affinity score between two nodes obtained from eqn (2) is symmetric. It is achieved by:

$$\tilde{W} = D^{-1/2} \times W \times D^{-1/2}, \qquad (3)$$

where W is the weighted adjacency matrix of the network (storing the edge weights w_{ij}), D is the diagonal matrix with the diagonal entries as $d_{ii} = \sum_j w_{ij}$. As such, $\tilde{w}_{ij} = w_{ij}/\sqrt{d_{ii} \times d_{jj}}$.

The restarting probability r takes the value from 0 to 1, controlling the range from the starting node s that the walker will explore. The higher the value of r, the more likely the walker is to visit the nodes centered on the starting node s. At the extreme when $r = 0$, the walker freely moves to the neighbors at each step without restarting, *i.e.*, following a random walk (RW).

We implemented a threaded version of RWR (*i.e.*, parallelized RWR using multiple threads) to pre-compute the affinity scores between all nodes, stored as affinity matrix A with the column s consisting of $\vec{p}_s(\infty)$. With this matrix, the affinity score to a set of nodes (*e.g.*, disease-related domains) was calculated using:

$$\vec{p}_N(\infty) = \frac{1}{n} \sum_{s \in N} \vec{p}_s(\infty), \qquad (4)$$

where N is a set of nodes being related, n for the size of N, and $\vec{p}_N(\infty)$ contains the affinity score of all nodes in the network to the set of nodes N. This is equivalent to the steady probability vector calculated according to eqn (2) but with the starting probability vector having equal probabilities (*i.e.*, $1/n$) for each node in N and 0 for the rest. As with $\vec{p}_s(\infty)$, the steady probability vector $\vec{p}_N(\infty)$ can also be considered as the "influential impact" over the network, but collectively imposed by a set of starting nodes N.

The motivation behind using eqn (2) to derive eqn (4) is that the pre-computed affinity matrix A can be extensively reused for obtaining affinity scores between any combinations of nodes according to eqn (4). For instance, such a strategy can dramatically relieve the computational burden involved in sampling a large number of random node combinations. Although the implementation of RWR in eqn (2) for the pair-wise affinity score does not scale for large networks (the runtime is quadratic to the number of nodes in the network), this pre-computation does permit some flexibility in the downstream use, in particular for statistical testing.

Performance evaluation using leave-one-out cross-validation

We conducted a leave-one-out cross-validation test to evaluate our ability to recover ontology terms annotated to domains. For every term, we performed one calculation for each domain annotated with that term, ignoring (leaving out) from the calculation the prior knowledge of that annotation. Each calculation involves a candidate set and a seed set. The candidate set consists of the domain in question, together with all the nodes (domains) in the network that are not annotated with the term in question. The seed set includes all of the remaining domains (*i.e.*, without the domain in question) annotated with the term in question. Based on eqn (4), we calculated the affinity score of each domain in the candidate set with respect to the seed set. The higher the ranking of the domain in question relative to the rest of the candidate set, the greater the ability to recover the annotation, and hence the greater the contribution to the overall value of the functional network. As a comparison to using eqn (4), we also calculated the affinity score only taking into account the direct neighbors (DN) of the seed set. That is, for each domain (from the candidate set) we summed up its neighboring edge weights if and only if its direct neighbors are within the seed set, and subsequently obtained the ranking of the domain in question. Unlike the ranking obtained using the global network information in eqn (4), the ranking based on DN uses only the local network information.

We used two performance measures to quantify how well the domains ranked relative to their candidate sets. One measure is the rank enrichment, which is calculated as the expected median rank (*i.e.*, median of candidate set size) divided by the rank of the domain in question. The reported rank enrichment is averaged over all domains for a term (*i.e.*, term-specific rank enrichment). Also, we considered together all terms within each ontology and reported the ontology-specific rank enrichment. The second alternative to the rank enrichment measure is the rank ROC curve for each term (also calculated for the whole ontology). By varying the rank threshold, the true positive rate *versus* the false positive rate can be plotted as a ROC curve. The area under the ROC curve (AUC) is used as the resultant quantitative measure, with an AUC of 1 indicating perfect performance and an AUC of 0.5 indicating random performance.

Building the connectivity matrix

The connectivity matrix was built by identifying statistically significant similarities between (non-exclusive) sets of domains, defined by association to an ontology term (for example, one set of domains annotated by a drug term, the other by a disease term). It was achieved using the pre-computed domain pair-wise affinity scores described above (being stored as affinity matrix A). Given two sets of domains, S_1 and S_2, we calculate a connectivity score (C-score; see eqn (5)) to quantify their similarity on the dcGOnet.

$$\text{C-score}(S_1, S_2) = \frac{\text{OBS}(S_1, S_2) - \mu(S_1, S_2)}{\sigma(S_1, S_2)}, \qquad (5)$$

Here, $\text{OBS}(S_1, S_2)$ is the observed affinity scores summed over all pair-wise combinations of domains from the two sets, S_1 and S_2. $\mu(S_1, S_2)$ and $\sigma(S_1, S_2)$ are respectively the mean and standard deviation estimated according to a null distribution. This null distribution was estimated using a novel randomization procedure, which respects the size of each set and node degree of its members. For a domain set S, we generate a random instance S^b, wherein domains randomly sampled from the network are the same in number as in S, and the degree distribution is also approximated by that of S. Such a degree distribution-approximating randomization is conceptually similar to a reference model previously reported by Erten *et al.*[26] Once a random instance has been sampled, we calculate the expected sum of affinity scores for this instance, denoted as $\text{EXP}^b(S_1, S_2)$, in a similar way to $\text{OBS}(S_1, S_2)$. By repeating this randomization procedure B times (2000 in this case), we obtain a sampling of the null distribution: $\text{EXP}^b(S_1, S_2)$, $b = 1, \ldots, B$. This null distribution is used to estimate the mean $\mu(S_1, S_2)$, the standard deviation $\sigma(S_1, S_2)$ and the empirical P-value $p(S_1, S_2)$, formulated as:

$$\mu(S_1, S_2) = \frac{1}{B}\sum_{b \in B} \text{EXP}^b(S_1, S_2), \qquad (6)$$

$$\sigma(S_1, S_2) = \frac{1}{B-1}\left\{\sum_{b \in B}\left[\text{EXP}^b(S_1, S_2) - \mu(S_1, S_2)\right]^2\right\}^{1/2}, \qquad (7)$$

$$p(S_1, S_2) = \frac{\sum_{b \in B} I\left\{\text{EXP}^b(S_1, S_2) \geq \text{OBS}(S_1, S_2)\right\}}{B}, \qquad (8)$$

where $I\{x\}$ is an indicator function that returns 1 if $x \geq 0$ and 0 if $x < 0$. Using the Benjamini–Hochberg derived step-up false discovery rate (FDR) procedure,[27] we correct P-values for multiple hypothesis testing. The connectivity matrix is represented by

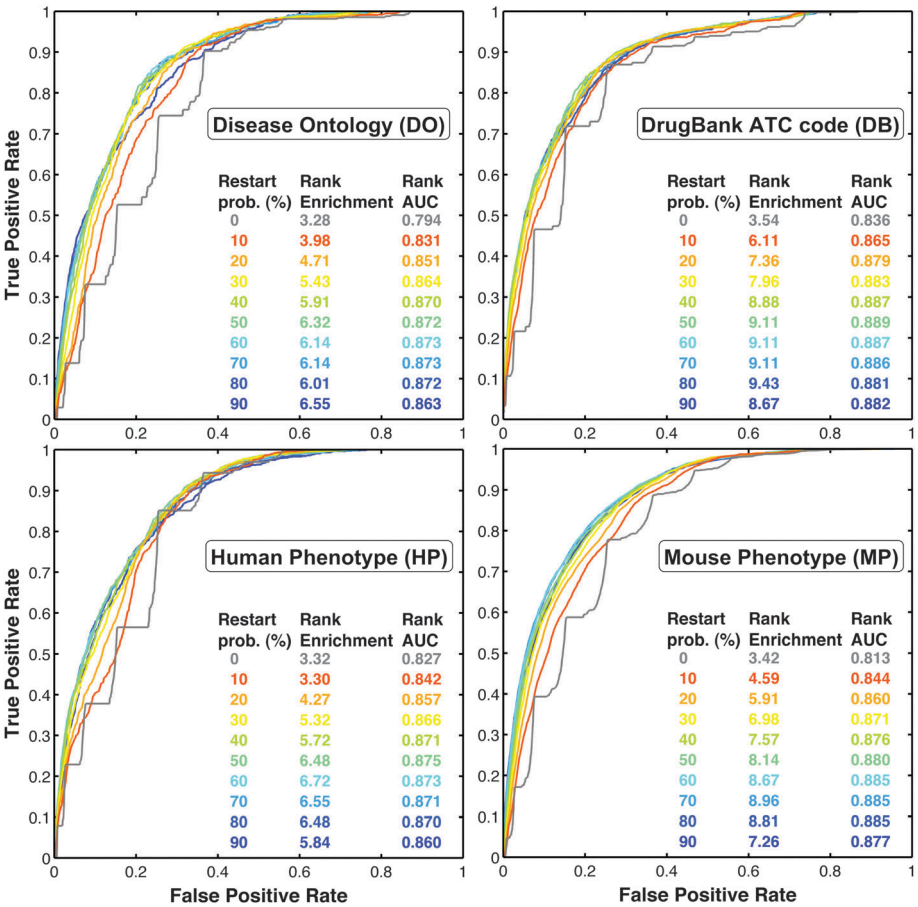

Fig. 1 Performance comparisons for RWR using different restarting probabilities. For each of the four ontologies, ROC curves compare the overall performance measure, along with the rank enrichment and rank AUC.

the C-scores between ontology terms. The higher the C-score is, the stronger the interconnection is. An FDR of 0.05 or lower is used to define statistical significance.

For comparison, we also implemented two conventional methods more commonly used for inferring the connectivity matrix. The first is the method based on the hyper-geometric distribution (HGD) of shared domain annotations. The presence of an interconnection between two terms is inferred if the observed overlap between their annotated domains is significantly more than would be expected by chance under the hyper-geometric distribution. The statistical significance is evaluated by Fisher's exact test, reporting a *P*-value (and FDR after accounting for multiple hypothesis tests). The second method is using direct neighbor information (DNI). Unlike using the pre-computed affinity matrix by the proposed RWR method, this method operates directly on the dcGOnet and calculates OBS(S_1, S_2) as the sum over all neighboring edge weights observed between the two sets of domains. The significance is estimated using the degree distribution-approximating randomization procedure as before.

Results and discussion

Walking on the dcGOnet globally predicts disease, drug and phenotype terms for domains

In this section, we examine the results of the leave-one-out cross-validation test (as measured by the rank enrichment and rank AUC; see Methods) on predicting ontology labels for domains. The ontologies tested include Disease Ontology (DO), DrugBank ATC codes (DB), Human Phenotype (HP) and Mouse Phenotype (MP) ontologies. For each ontology, we evaluate: (i) the method using global network information (RWR with different restarting probabilities and RW); (ii) the method using local network information (direct neighbors, DN); (iii) the effect of network size on the performance; and (iv) the effect of using different parts of GO to construct the network used by our method.

A key result from these performance evaluations is that utilising the network in a global manner (by using RWR) is significantly better than using local network information. The remarkable performance achieved by RWR is insensitive to both the choice of parameters and the size of the network used. Another important result from the experiment is that it is optimal to include all possible GO terms when constructing the functional domain network; this was not expected as we and others had recently observed that domains are better described by the molecular function (MF) part of GO than the biological process (BP) or cellular component (CC) parts of GO.[28,29] Below we describe these results in detail.

Using RWR is insensitive to the choice of restarting probability. The restarting probability in eqn (2) controls the extent of influence that the starting node (or a set of starting nodes) exerts on the network topology. Fig. 1 shows that the effect on the curve of changing the restart probability is relatively minor for most of the range between 0.1 and 0.9. This conclusion holds for all four ontologies being tested. To simplify the presentation of results for the reader, we selected 0.5 as the default value, and use this value when referring to RWR hereinafter.

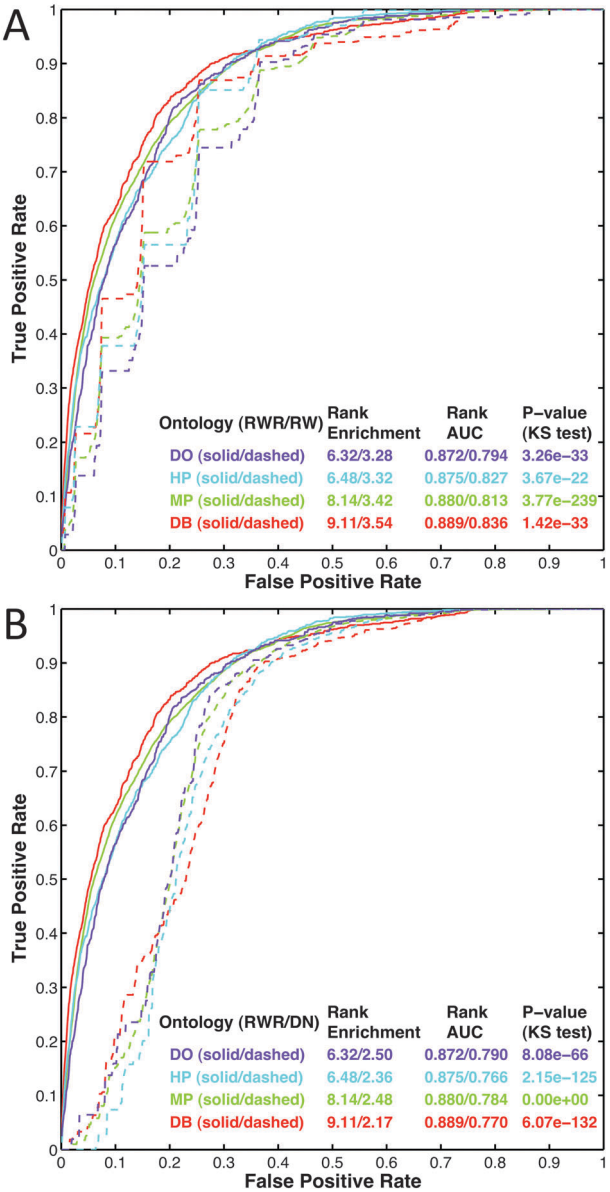

Fig. 2 Performance comparisons of RWR against RW and the method using direct neighbors (DN). (A) RWR against RW. (B) RWR against DN. In addition to the rank enrichment and rank AUC, *P*-value based on two-sample *KS*-test is also reported to assess the significance of the performance difference.

Using RWR is significantly better than using RW. A special case of RWR without restart is RW, which is not influenced by the start node. In Fig. 1, it is clear that the performance of RW (*i.e.*, RWR with 0 restarting probability) on this test is inferior to that of RWR. To show whether or not the difference is statistically significant, we applied the two-sample Kolmogorov–Smirnov (KS) test to compare the two distributions of rank values: one resulting from RWR, and the other from RW. According to the reported *P*-value, the difference between RWR and RW is statistically significant for each of four ontologies (Fig. 2A).

Using global network information is significantly better than using local network information. As a comparison to methods that use local network information, we also calculated

the rankings using direct neighbors (DN; see Methods) to the seed set. As before, both the rank enrichment and rank AUC were calculated. Compared to DN (Fig. 2B), the performance of RWR is confirmed to be significantly better.

The RWR method is robust to network size. To investigate the effect of network size on the performance of RWR, we also created networks of different sizes by selecting various fractions of the top interactions (and their associated weights as well) with the highest weights. As illustrated in Fig. 3A, the performance is almost unchanged when using 40% or more of the top interactions. Although the results appear to become slightly poorer as the size of the network decreases, the KS test did not reveal the difference to be statistically significant. Thanks to this robustness of the RWR method, we are able to reduce the network size without loss of performance. For the dcGOnet, partly to lessen the computational burden of any operations on the network, we include only the top 50% of interactions; this threshold corresponds to those interactions with a semantic similarity value of 0.4 or higher.

Using all possible GO annotations is better than only using molecular function. In dcGO domains are functionally annotated with all three complementary sub-ontologies (*i.e.*, MF, BP, CC) from GO. Since MF best describes protein domains,[28,29] for comparison we constructed a domain network using only annotations from the MF sub-ontology. The results clearly show that including the other GO sub-ontologies in addition to MF can seriously enhance the performance of the method (Fig. 3A *versus* Fig. 3B). This observation suggests that the connections between domains in the constructed dcGOnet capture a broader context than the functions of the domains themselves; including the BP and CC sub-ontologies increases the predictive power.

Performance is not limited to specific aspects of knowledge. In addition to the evaluation of aggregate performance, we also looked at the performance for individual ontology terms. In Fig. 4 we list those terms with an AUC better than that for the overall ontology. These above-average terms cover a broad range of knowledge, showing the generality of the method. As a matter of fact, the below-average and poorly ranked terms are also not knowledge-specific (data not shown).

Applying dcGOnet to build a disease-drug-phenotype connectivity matrix

The aforementioned power of the functional domain network (in predicting disease, drug and phenotype terms associated with domains) reflects the inherent relationship between domain

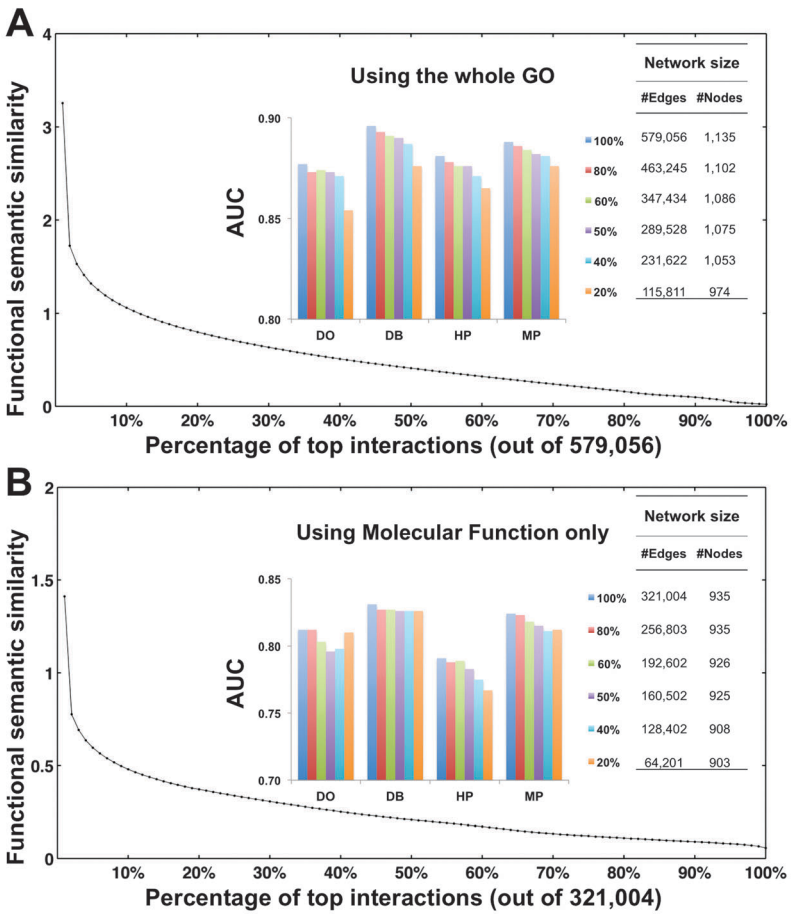

Fig. 3 Performance comparisons using functional domain networks from different sources and the networks of different sizes. The curve displays functional semantic similarity for varying percentage of top interactions. The bar graph inserted reports the performance for different network sizes. (A) Functional domain network constructed by using the whole GO. (B) Functional domain network constructed by using MF sub-ontology only.

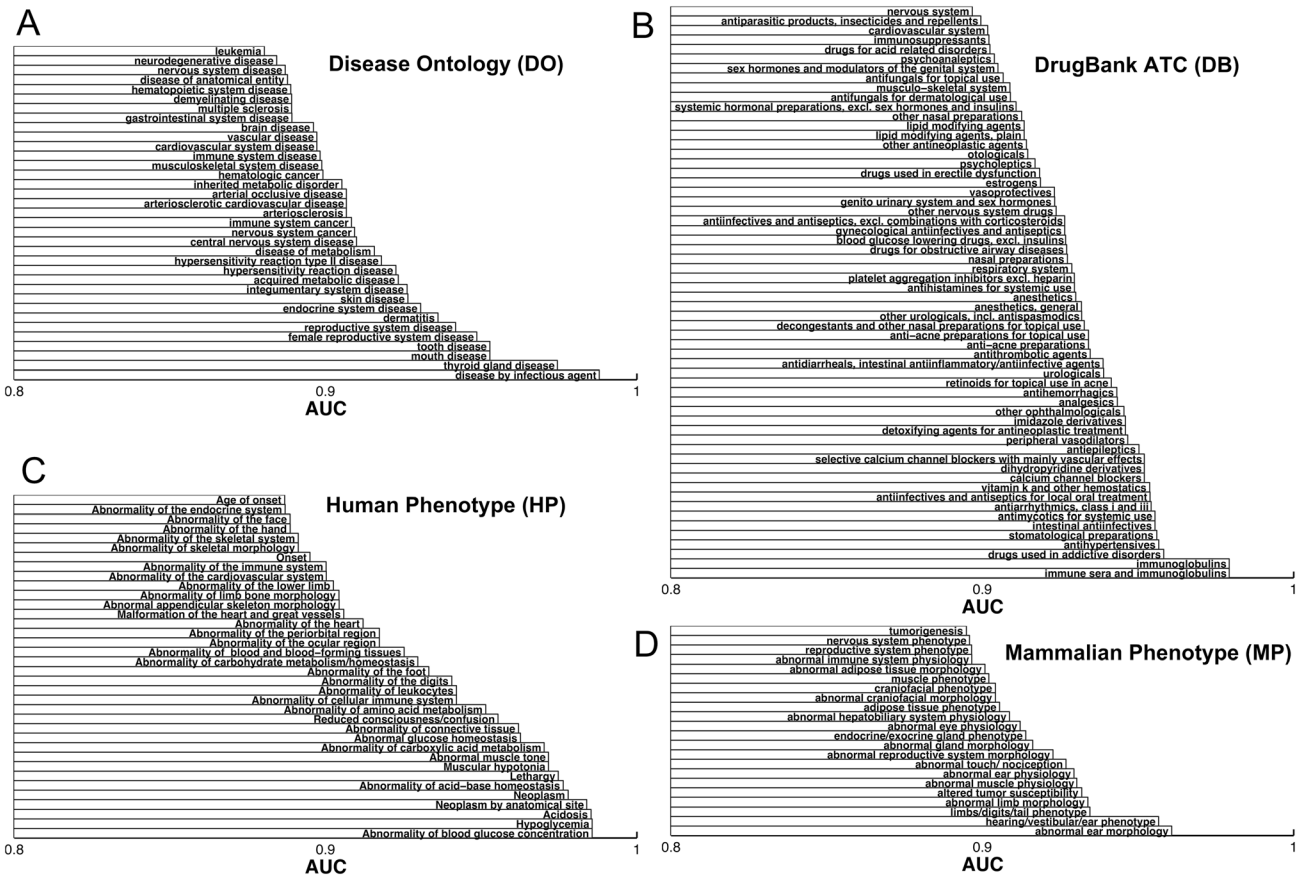

Fig. 4 Ontology term-specific performance comparisons.

function and disease (or phenotypes/drugs). Therefore, we further explored this domain network by examining the possibility of revealing inter-ontology relationships. To do so, we developed a novel method for building a connectivity matrix (see Methods for details). Briefly, we devised a connectivity score to define the global correspondence between two sets of domains within dcGOnet. The connectivity score (C-score) was calculated from domain pair-wise affinity scores pre-computed by RWR, and was statistically adjusted/normalized using a degree-approximating randomization procedure. As a result, the connectivity matrix not only reveals the strength of an interconnection, but also gives the statistical significance of it. To help visualize the disease-drug-phenotype connectivity matrix, we applied two-dimensional hierarchal clustering to reorder terms (but still constrained within the ontology of their origin).

We also compared the RWR method with two simpler methods: one based on hyper-geometric distribution (HGD) and the other using direct neighbor information (DNI). The HGD method does not use the network and only requires the overlap of domain annotations; this method cannot reveal any relationships where there is no or trivial intersection observed. The DNI method does utilize the information in the network, but this method cannot make any inference if there is no direct connection. In both cases, however, the RWR method may still reveal the hidden interconnection by exploring the functional similarities between domains in a global manner. Indeed at the same false discovery rate (FDR < 0.05), the RWR method almost doubles the interconnections that are commonly identified by both RWR and HGD methods (Fig. S1A, ESI†). Similar results were also observed when comparing against the DNI method (Fig. S2A, ESI†). These comparisons establish the necessity of going through the functional network construction and the RWR procedures rather than applying simpler methods. The lack of a gold-standard benchmark means that it is not possible to systematically quantify the accuracy of each of the methods being compared. We are able however to find some indications of the superiority of the RWR method *versus* the simpler methods. Firstly we find that many of the top predictions by the RWR method have empirical support, yet are missed by the others, in particular for drug-disease connections (Tables 1–3; see the rest of this section). Furthermore, examining the distribution of FDR values (Fig. S1B and S2B, ESI†) reveals the following observations: (i) predictions common to the comparing methods tend to have lower FDR values (more significant) as these are easy targets; (ii) predictions unique to the HGD or DNI method tend to have higher FDR values, and hence more expected false positives than for the common predictions; (iii) in contrast, unique predictions by the RWR method, although sharing in this trend, do so to a much

Table 1 Top drug-disease connections by RWR

DrugBank ATC code			
ID	Name	Predicted connections to diseases (rank[a])	External evidence supporting the prediction
B01AC	Platelet aggregation inhibitors excl. heparin	Brain disease (1st); central nervous system disease (2nd); thyroid gland disease (6th)[b]	Drugs in this group, *e.g.*, aspirin, are extensively used as an analgesic[39] that acts in various ways on the peripheral and central nervous systems. Use aspirin to reduce the inflammation for hyperthyroidism.
A11HA03	Tocopherol (vitamin E)	Hepatobiliary disease (5th)[b]; leukemia (20th)[b]	A case has been reported using vitamin E to treat vitamin E-deficient neuropathy in a patient with chronic hepatobiliary disease.[34] Tocopherol has an *in vivo* antileukemic action.[30]
V03AB	Antidotes	Brain disease (13th); central nervous system disease (14th)[b]; hepatobiliary disease (18th)[b]	Sugammadex in this group is indicated for reversal of neuromuscular blockade induced by rocuronium or vecuronium.[36] Moreover, patients with hepatobiliary disease have prolonged blockade by rocuronium.[35]
D10AD	Retinoids for topical use in acne	Prostate cancer (10th)[b]; male reproductive organ cancer (11th); female reproductive organ cancer (15th); skin disease (16th)[b]; reproductive organ cancer (17th)[b]	Isotretinoin in this group is used for preventing skin cancer,[31] for treating recurrent prostate cancer[32] (male reproductive organ cancer). Retinoids have potential chemopreventive and therapeutic roles in cervical cancer[33] (female reproductive organ cancer).

[a] The top 25 predictions are available in Table S1 (ESI). [b] Those connections missed out by the other two methods.

Table 2 Top disease-phenotype connections in human by RWR

Disease Ontology		
ID	Name	Predicted connections to human phenotypes (rank[a])
DOID:50	Thyroid gland disease	Abnormal interferon secretion (1st); abnormal interferon-gamma secretion (2nd)
DOID:229	Female reproductive system disease	Abnormal limb long bone morphology (3rd); abnormal limb bone morphology (4th); abnormal limb morphology (5th)[b]; limbs/digits/tail phenotype (7th); abnormal long bone morphology (21st)
DOID:0050117	Disease by infectious agent	Abnormal Peyer's patch morphology (15th); abnormal gut-associated lymphoid tissue morphology (16th)
DOID:0060056	Hypersensitivity reaction disease	Abnormal inflammatory response (24th)[b]; increased inflammatory response (25th)

[a] The top 25 predictions are available in Table S2 (ESI). [b] Those connections missed out by the other two methods.

Table 3 Top cross-species phenotype connections by RWR

Human Phenotype		
ID	Name	Predicted connections to mouse phenotypes (rank[a])
HP:0000271	Abnormality of the face	Abnormal facial morphology (4th); limbs/digits/tail phenotype (19th)
HP:0002795	Functional respiratory abnormality	Abnormal emotion/affect behavior (7th); abnormal behavioral response to xenobiotic (12); abnormal eyelid morphology (23)
HP:0001760	Abnormality of the foot	Abnormal digit morphology (13th); abnormal autopod morphology (14th)
HP:0002715	Abnormality of the immune system	Abnormal immune tolerance (16th)[b]; abnormal self tolerance (17th); autoimmune response (18th)

[a] The top 25 predictions are available in Table S3 (ESI). [b] Those connections missed out by the other two methods.

lesser extent and are therefore of higher quality (with fewer expected false positives).

From the disease-drug-phenotype connectivity matrix obtained using the RWR method, we display in Fig. 5 those terms with above-average AUC (as listed in Fig. 4). For reference, the matrix for below-average AUC terms is also provided in Fig. S3 (ESI†). The matrix reveals an uneven distribution of connectivity, in particular for the interconnections between disease terms and drug terms (Fig. 5). Most of the drug terms have statistically significant connections to nervous system/brain diseases and metabolism diseases. The promiscuous nature of these diseases might reflect the current practices of drug development in combating them. The matrix recovers many well-known connections, yet also suggests many new ones. Tables S1–S3 (ESI†) list the top 25 predictions. Below we provide examples to illustrate the biological relevance of the predicted connections.

Drug-disease connections. Most of the top predictions (Table 1) are consistent with existing knowledge: the first prediction connects platelet aggregation inhibitors (*e.g.*, aspirin) with brain and central nervous system diseases; tocopherol (vitamin E), exerting antileukemic activity,[30] is successfully predicted as linked to leukemia; antidotes are connected to brain and central nervous system diseases; retinoids are predicted as connected with skin disease,[31] prostate cancer,[32] and cervical cancer.[33] We also uncover new connections, some of

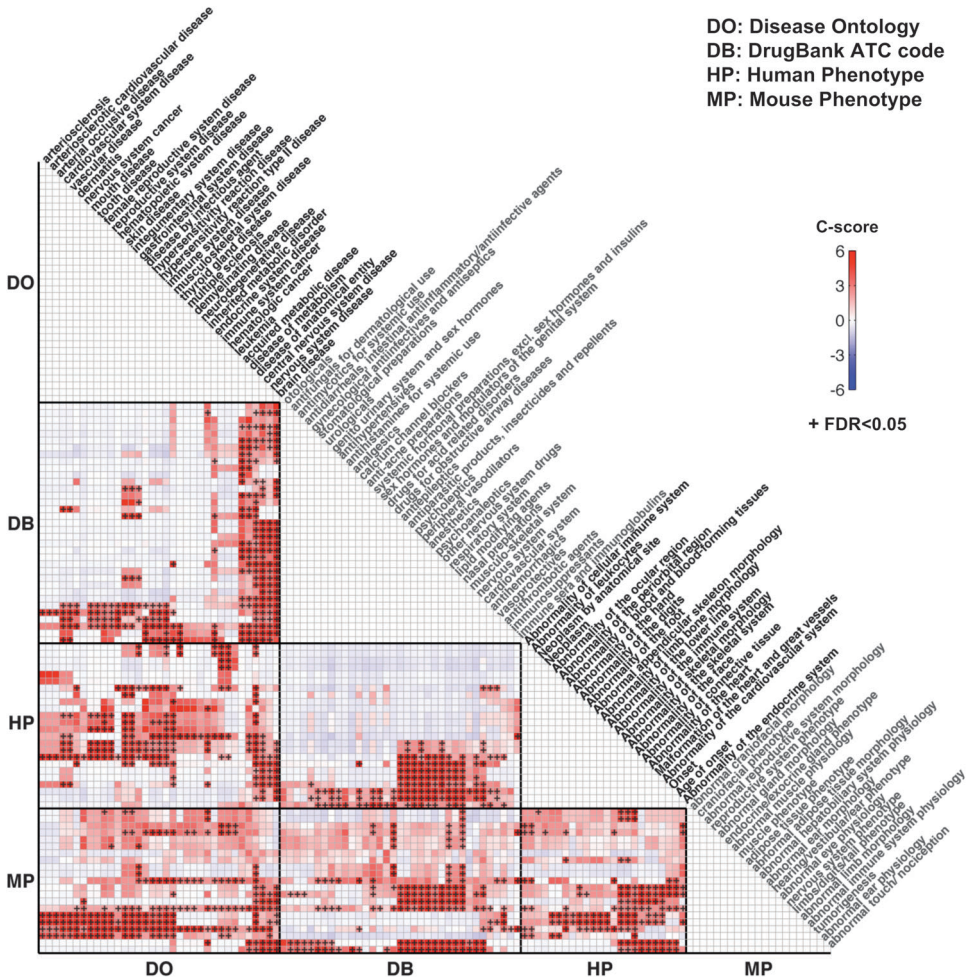

Fig. 5 A disease-drug-phenotype connectivity matrix. The matrix is colour-encoded according to connectivity scores, and is also marked with the statistically significant connections.

which have been implicated in the literature. For example, tocopherol is predicted to link with hepatobiliary disease, for which we found supporting evidence in a clinical case report.[34] Our predictions also suggest that antidotes are related to hepatobiliary disease; interestingly, patients with hepatobiliary disease exhibit prolonged blockade by rocuronium,[35] which can be reversed by an antidote, Sugammadex.[36] Drug repositioning[37] is a field dedicated to find new indications/diseases for existing drugs. The predicted new connections for tocopherol and antidotes suggest that our approach is a viable strategy for suggesting drug repositioning targets.

Disease-phenotype connections in human. As illustrated in Table 2, there is a predicted connection between thyroid gland disease and abnormal interferon secretion phenotype. This is as expected since long-term treatment with interferon could lead to thyroid disorder.[38] Our analysis links abnormal lymphoid tissue and Peyer's patches (*i.e.*, lymphoid nodules) to 'infectious disease' and 'abnormal/increased inflammatory response to hypersensitivity reaction disease'; both are in agreement with accepted knowledge. Surprisingly, we suggest a strong link between abnormal limb bone morphology and female reproductive system disease, for which we have not yet found an explanation.

Cross-species phenotype connections. Cross-species phenotype comparisons are extremely important for research since much research hinges on the principle that knowledge obtained from model organisms can be transferred to human. Our predictions on phenotype connections between human and mouse shown in Table 3 are mostly self-explanatory. Beyond the two organisms mentioned here (mouse and human), we have also built full connections between all organisms for which phenotype ontologies are available using dcGO (see ESI† website at http://supfam.org/SUPERFAMILY/dcGO/dcGOnet.html).

Conclusion

A global random walk on a functional domain network is able to recover labels for disease, drug or phenotype-related ontologies. These (biomedical) ontologies are independent of the (functional) ontology from which the network was derived, supporting the hypothesis that functionally related domains tend to impact the same drugs/diseases/phenotypes as each other.

Moreover, globally exploring the functional domain network successfully identifies relationships between drugs, diseases and phenotype ontology terms, supporting the use of protein domains as mediators to bridge between ontologies of different types. In this paper we describe the formulation, implementation and application of our strategy. The network-based approach developed for building our matrix of relationships is general, and could additionally be applied at the protein level. Combining network biology with ontology annotations both at the domain level however, allows for the extraction of cross-knowledge and cross-species relationships that are often obscured from a simple protein level view. As a product of this work, the functional domain network itself has further potential as a reference, *e.g.*, for studying the co-occurrence of context-dependent networks in genomes.

Abbreviations

AUC	Area under the ROC curve
BP	Biological process
C-score	Connectivity score
CC	Cellular component
DB	DrugBank ATC codes
dcGO	Domain-centric Gene Ontology
DN	Direct neighbors
DNI	Direct neighbor information
DO	Disease Ontology
FDR	False discovery rate
GO	Gene Ontology
HGD	Hyper-geometric distribution
HP	Human Phenotype
IC	Information content
MF	Molecular function
MICA	Most Informative Common Ancestor
MP	Mouse Phenotype
ROC	Receiver operating characteristic
RW	Random walk
RWR	Random walk with restart
SCOP	Structural Classification of Proteins

Acknowledgements

We thank anonymous referees for valuable suggestions. This work was supported by the Biotechnology and Biological Sciences Research Council [grant number BB/G022771/1 to J.G.].

References

1 A. L. Barabasi and Z. N. Oltvai, *Nat. Rev. Genet.*, 2004, **5**, 101–113.
2 U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach and E. E. Wanker, *Cell*, 2005, **122**, 957–968.
3 J. D. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth and M. Vidal, *Nature*, 2004, **430**, 88–93.
4 J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth and M. Vidal, *Nature*, 2005, **437**, 1173–1178.
5 L. Hakes, J. W. Pinney, D. L. Robertson and S. C. Lovell, *Nat. Biotechnol.*, 2008, **26**, 69–72.
6 G. T. Hart, A. K. Ramani and E. M. Marcotte, *Genome Biol.*, 2006, **7**, 120.
7 I. Lee, B. Lehner, C. Crombie, W. Wong, A. G. Fraser and E. M. Marcotte, *Nat. Genet.*, 2008, **40**, 181–188.
8 I. Lee, B. Lehner, T. Vavouri, J. Shin, A. G. Fraser and E. M. Marcotte, *Genome Res.*, 2010, **20**, 1143–1153.
9 A. L. Barabasi, N. Gulbahce and J. Loscalzo, *Nat. Rev. Genet.*, 2011, **12**, 56–68.
10 A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia and A. G. Murzin, *Nucleic Acids Res.*, 2008, **36**, D419–D425.
11 S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari and R. Jothi, *Nucleic Acids Res.*, 2011, **39**, D730–D735.
12 S. Wuchty and E. Almaas, *BMC Evol. Biol.*, 2005, **5**, 24.
13 Z. Wang, X. C. Zhang, M. H. Le, D. Xu, G. Stacey and J. Cheng, *PLoS One*, 2011, **6**, e17906.
14 H. Fang and J. Gough, *Nucleic Acids Res.*, 2013, **41**, D536–D544.
15 C. Pesquita, D. Faria, A. O. Falcao, P. Lord and F. M. Couto, *PLoS Comput. Biol.*, 2009, **5**, e1000443.
16 L. Grady, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, 1768–1783.
17 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nat. Genet.*, 2000, **25**, 25–29.
18 B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel and S. Lewis, *Nat. Biotechnol.*, 2007, **25**, 1251–1255.
19 D. A. de Lima Morais, H. Fang, O. J. Rackham, D. Wilson, R. Pethica, C. Chothia and J. Gough, *Nucleic Acids Res.*, 2011, **39**, D427–D434.
20 L. M. Schriml, C. Arze, S. Nadendla, Y. W. Chang, M. Mazaitis, V. Felix, G. Feng and W. A. Kibbe, *Nucleic Acids Res.*, 2012, **40**, D940–D946.
21 P. N. Robinson, S. Kohler, S. Bauer, D. Seelow, D. Horn and S. Mundlos, *Am. J. Hum. Genet.*, 2008, **83**, 610–615.

22 C. L. Smith and J. T. Eppig, *Wiley Interdiscip. Rev.: Syst. Biol. Med.*, 2009, **1**, 390–399.
23 C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo and D. S. Wishart, *Nucleic Acids Res.*, 2011, **39**, D1035–D1041.
24 S. Brin and L. Page, *Comput. Networks ISDN*, 1998, **30**, 107–117.
25 H. Tong, C. Faloutsos and J.-Y. Pan, in *Proceedings of the Sixth International Conference on Data Mining*, IEEE Computer Society, 2006.
26 S. Erten, G. Bebek, R. M. Ewing and M. Koyuturk, *BioData Min.*, 2011, **4**, 19.
27 Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B*, 1995, **57**, 289–300.
28 H. Fang and J. Gough, *BMC Bioinf.*, 2013, **14**, S9.
29 P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Toronen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D. W. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta, S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kassner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Honigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Bjorne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. Sternberg, N. Skunca, F. Supek, M. Bosnjak, P. Panov, S. Dzeroski, T. Smuc, Y. A. Kourmpetis, A. D. van Dijk, C. J. Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner, C. Orengo, B. Rost, S. D. Mooney and I. Friedberg, *Nat. Methods*, 2013, **10**, 221–227.
30 G. A. dos Santos, R. S. Abreu e Lima, C. R. Pestana, A. S. Lima, P. S. Scheucher, C. H. Thome, H. L. Gimenes-Teixeira, B. A. Santana-Lemos, A. R. Lucena-Araujo, F. P. Rodrigues, R. Nasr, S. A. Uyemura, R. P. Falcao, H. de The, P. P. Pandolfi, C. Curti and E. M. Rego, *Leukemia*, 2011, **26**, 451–460.
31 N. Levine, T. E. Moon, B. Cartmel, J. L. Bangert, S. Rodney, Q. Dong, Y. M. Peng and D. S. Alberts, *Cancer Epidemiol., Biomarkers Prev.*, 1997, **6**, 957–961.
32 M. Shalev, T. C. Thompson, A. Frolov, S. M. Lippman, W. K. Hong, H. Fritsche and D. Kadmon, *Clin. Cancer Res.*, 2000, **6**, 3845–3849.
33 J. Abu, M. Batuwangala, K. Herbert and P. Symonds, *Lancet Oncol.*, 2005, **6**, 712–720.
34 H. Y. Ko and I. Park-Ko, *Arch. Phys. Med. Rehabil.*, 1999, **80**, 964–967.
35 M. M. van Miert, N. B. Eastwood, A. H. Boyd, C. J. Parker and J. M. Hunter, *Br. J. Clin. Pharmacol.*, 1997, **44**, 139–144.
36 O. Sacan, P. F. White, B. Tufanogullari and K. Klein, *Anesth. Analg.*, 2007, **104**, 569–574.
37 J. T. Dudley, T. Deshpande and A. J. Butte, *Briefings Bioinf.*, 2011, **12**, 303–311.
38 M. Gelu-Simeon, A. Burlaud, J. Young, G. Pelletier and C. Buffet, *World J. Gastroenterol.*, 2009, **15**, 328–333.
39 Z. Gaciong, *Thromb. Res.*, 2003, **110**, 361–364.