Novel search method for the discovery of functional relationships

Fidel Ramirez, Glenn Lawyer and Mario Albrecht*
Max Planck Institute for Informatics, Campus E1.4, 66123 Saarbrücken, Germany

ABSTRACT

Motivation: Numerous annotations are available that functionally characterize genes and proteins with regard to molecular process, cellular localization, tissue expression, protein domain composition, protein interaction, disease association and other properties. Searching this steadily growing amount of information can lead to the discovery of new biological relationships between genes and proteins. To facilitate the searches, methods are required that measure the annotation similarity of genes and proteins. However, most current similarity methods are focused only on annotations from the Gene Ontology (GO) and do not take other annotation sources into account.

Results: We introduce the new method BioSim that incorporates multiple sources of annotations to quantify the functional similarity of genes and proteins. We compared the performance of our method with four other well-known methods adapted to use multiple annotation sources. We evaluated the methods by searching for known functional relationships using annotations based only on GO or on our large data warehouse BioMyn. This warehouse integrates many diverse annotation sources of human genes and proteins. We observed that the search performance improved substantially for almost all methods when multiple annotation sources were included. In particular, our method outperformed the other methods in terms of recall and average precision.

Contact: mario.albrecht@mpi-inf.mpg.de

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on May 16, 2011; revised on October 17, 2011; accepted on November 9, 2011

1 INTRODUCTION

Similarity search plays an important role in biological, pharmaceutical and medical investigations. For instance, the introduction of the BLAST algorithm by Altschul et al. (1990) to search for similar sequences has been considered a milestone in genomics (Bahcall, 2007). Other similarity search methods to mine databases of 3D molecule conformations have been important for drug discovery (Willett et al., 1998). In addition, the growing availability of annotations that characterize genes and proteins (Reeves et al., 2008) opens the new possibility to find biological relationships by similarity searches based on function, domain composition, disease association, tissue expression, etc. For example, the identification of similarly annotated genes and proteins can reveal new gene–disease associations (Aerts et al., 2006), suggest novel protein functions (Fridberg, 2006) and indicate new drug targets (Chan et al., 2010).

In general, similarity searches compute pairwise similarities of a query with the entities in a database to obtain a ranked list of high-scoring similarities. In particular, a number of methods have been proposed for the quantification of pairwise similarities of gene and protein annotations. Most of those functional similarity methods are based on Gene Ontology (GO) annotations (Benabderrahmane et al., 2010; Chabalier et al., 2007; del Pozo et al., 2008; Lerman and Shakhnovich, 2007; Lord et al., 2003; Mistry and Pavlidis, 2008; Pesquita et al., 2008; Popescu et al., 2006; Schlicker et al., 2006; Sevilla et al., 2005; Speer et al., 2004). However, the last years have shown a dramatic growth in datasets that result from high-throughput experiments and computational work and yield annotation sources that provide manifold information about, for instance, protein interactions, signaling circuits, metabolic pathways, cellular localization, tissue expression, disease associations and protein domain architecture. Currently, only one similarity search method explicitly takes multiple annotation sources into account, namely, the kappa coefficient used by the DAVID Gene Functional Classification Tool (Huang et al., 2007). In contrast, the integration of multiple annotation sources into a network structure is often applied in the context of gene function prediction (Huttenhower et al., 2009; Wang and Marcotte, 2010; Warde-Farley et al., 2010).

When developing efficient methods for searching through gene and protein annotation data, a particular task is the construction of data structures that represent the annotations. Most methods rely on the graph structure of GO to estimate quantitative semantic relationships among the gene/protein annotations (Pesquita et al., 2009). However, the GO structure limits the inclusion of non-ontological (i.e. non-GO) annotations into methods. A flattened representation of the GO hierarchy solves this problem and allows the use of multiple annotation sources (Pesquita et al., 2006). This representation implicitly contains the ontological relationships and allows the inclusion of non-ontological annotations as part of the array. This avoids the inference of relationships through the hierarchical structure of GO. GO-based similarity methods that use this data structure are COS (Chabalier et al., 2007), simGIC (Pesquita et al., 2008) and TO (Mistry and Pavlidis, 2008). Although these methods do not consider annotation sources other than GO, they achieve better performance than methods such as those by Restnik (1999) and Lin (1998) that depend on the GO graph structure.

In the following, we will introduce the new method BioSim for similarity searches based on diverse annotation sources of gene and protein function and extend the existing methods cosine similarity, kappa coefficient, simGIC and TO to utilize annotations not only...
from GO, but also from 22 major biological databases for human genes and proteins. We will also compare the performance of BioSim with the other methods in different benchmarks.

2 MATERIALS AND METHODS

2.1 Annotation sources

Twenty-two publicly available annotation sources for human genes and proteins were integrated into our data warehouse BioMyn. These include functional annotations from all three GO categories (MF, molecular function; BP, biological process; CC, cellular component) (Camon et al., 2004) and from the UniProtKB controlled vocabulary of keywords (2008) and InterPro (Hunter et al., 2009); metabolic and signaling pathways from HumanCyc (Romero et al., 2005), KEGG (Kanehisa et al., 2008), and Reactome (Matthews et al., 2009); protein-protein interactions and protein complexes from CORUM (Ruepp et al., 2008), DIP (Salwinski et al., 2004), HiMAP (Rhodes et al., 2005), HPRD (Prasad et al., 2009), IntAct (Kerrien et al., 2007), MINT (Chau-Ayermonts et al., 2007), PDB (Bernstein et al., 2003; Velankar et al., 2005) and STRING (Jensen et al., 2009); disease associations from OMIM (Amberger et al., 2006); enzyme classifications from the Enzyme nomenclature database (Bairoch, 2000); gene expression data for different tissues and cell lines from the Novartis Gene Atlas (Su et al., 2002); Mammalian Phenotype Ontology annotations of human genes as provided by the Mouse Genome Database (Blake et al., 2011); and orthologs of protein sequences from OrthoMCL (Chen et al., 2006).

From the annotation sources, the functionally relevant features associated with individual genes and proteins were extracted. In the following, we refer to these features as annotation terms, which correspond, for example, to the specific molecular function (e.g. oxidoreductase activity) or domain (e.g. SH2) or pathway (e.g. glycolysis) annotated to genes and proteins. The different gene and protein identifiers used in the annotation sources were unified by mapping them to Entrez Gene ID and UniProtKB accession numbers. In total, our data warehouse contains 245,596 human Entrez Gene entries and 70,767 human UniProtKB protein entries (including 20,177 manually reviewed proteins in UniProtKB release 15.5, see Supplementary Material for further details). To enable comparisons between functional similarity methods using multiple annotation sources and those using only GO annotations, proteins with no available GO annotation were excluded. This resulted in a list of 18,076 protein entries out of 20,177 manually reviewed proteins in UniProtKB release 15.5.

2.2 Functional similarity methods

In the following, $A_X$ and $A_Y$ denote the sets of annotation terms associated with the gene products $X$ and $Y$, respectively. Annotations available for genes are transferred to the encoded proteins.

**BioSim:** Our novel functional similarity method BioSim is defined as follows:

$$\text{BioSim}(X, Y) = \prod_{t \in A_X \cap A_Y} p(t)$$

Here, $t \in (A_X \cap A_Y)$ is the set of annotation terms shared by $X$ and $Y$, and $p(t)$ is the probability that both $A_X$ and $A_Y$ contain the same term $t$ by chance. Since BioSim is the product of the probabilities $p(t)$, a score of zero represents the highest similarity and a score of one the lowest. This is in contrast to other methods described below, except TO. The probability $p(t)$ is estimated using the cumulative hypergeometric distribution:

$$p(t) = \frac{D \times (N_E - N_t)}{N_t \times (N_E - N_t - D)}$$

$N_t$ is the number of proteins in our database, and $N_E$ is the number of proteins annotated with term $t$. $D$ is the sum of $|A_X|$ and $|A_Y|$, that is, the total number of annotation terms for $X$ and $Y$. Therefore, the resulting probability $p(t)$ depends not only on the frequency $N_t$, and thus on the specificity, of the annotation term $t$, but also on $D$. This is an important property of BioSim and accounts for the annotation bias of intensively studied genes and proteins. A pair of proteins associated with large $D$ has an increased probability $p(t)$ to share the annotation term $t$ (i.e. a decreased functional similarity) in comparison to a pair of proteins associated with few annotations terms (small $D$).

**Term overlap length (TO):** TO represents the number of annotation terms shared by two proteins $X$ and $Y$ (Mistry and Pavlidis, 2008):

$$\text{TO}(X, Y) = |A_X \cap A_Y|$$

**Kappa coefficient (KC):** This method is used in the well-known DAVID Gene Functional Classification Tool (Huang et al., 2007). It computes a normalized difference of the observed number of annotation terms $O(X, Y)$ shared by two proteins $X$ and $Y$, and the expected number $E(X, Y)$ of shared annotation terms that are randomly chosen (Huang et al., 2007). It is defined as follows:

$$\text{KC}(X, Y) = \frac{O(X, Y) - E(X, Y)}{1 - E(X, Y)}$$

In the following, we describe the two methods simGIC and COS. Unlike the previous methods, both methods incorporate term weights based on the information content (IC) of a term $t$ (Resnik, 1995):

$$\text{IC}(t) = -\log N_t$$

Here, $N_t$ is the number of proteins annotated with term $t$ and $N$ the total number of proteins in our study.

**simGIC:** This method introduced in Pesquita et al. (2008) includes the summed information contents of shared versus all annotated terms for two proteins $X$ and $Y$:

$$\text{simGIC}(X, Y) = \sum_{t \in A_X \cap A_Y} \text{IC}(t)$$

**Cosine similarity (COS):** This classical method is defined as follows (Salton et al., 1975):

$$\text{COS}(X, Y) = \frac{\sum_{t \in A_X \cap A_Y} \text{IC}(t)}{\sum_{t \in A_X} \text{IC}(t)}$$

Here, $A_X$ and $A_Y$ are the annotation vectors of two proteins $X$ and $Y$, respectively. In each vector, the absence of an annotation term is represented by 0 and the presence by IC(t). This method was first used in the context of functional similarity by Chabali et al. (2007).

2.3 Evaluation methods

**Gold standard:** To evaluate the performance of the functional similarity methods, we collected a gold standard dataset composed of groups of proteins that are assumed to be functionally related (to a certain extent) and contained in the list of 18,076 proteins with at least one available GO annotation (as described above). The protein groups in the dataset were obtained from four benchmark categories that we limited to at most 400 groups per category: (i) 400 groups from protein complexes (selected randomly from a total of 2030 complexes in CORUM); (ii) 88 groups from sequence clusters of related protein sequences based on UniRef90 clusters (sequences of at least 90% identity) and thus with putatively similar functions; (iii) 355 groups from reliable protein-protein interactions (here, an interaction is regarded as reliable if it is reported in at least three different publications); and (iv) 400 groups from metabolic and signaling pathways (selected randomly from a total of 424 pathways in KEGG and Reactome). Groups of more than 90% identity) and thus with putatively similar functions; (iii) 355 groups from reliable protein–protein interactions (here, an interaction is regarded as reliable if it is reported in at least three different publications); and (iv) 400 groups from metabolic and signaling pathways (selected randomly from a total of 424 pathways in KEGG and Reactome). Groups of more than 20 proteins were excluded as being too general. The average group size was 6.7 proteins and the overall standard deviation 4.3. In total, the gold standard consisted of 1243 groups that covered 8130 proteins overall (some proteins were shared by different groups). In the following, we will refer to those groups as validation groups.
3 RESULTS AND DISCUSSION

3.1 Evaluating the performance of functional similarity methods

The performance of BioSim in identifying known functional similarities was compared with that of four other methods: TO, KC, simGIC and COS. The results were averaged over all validation groups. While all methods showed similar performance when using only GO annotations, the performance was improved when considering multiple annotation sources (Fig. 1). Notably, BioSim achieved better performance than the other methods. For instance, the top 20 hits of BioSim had an average recall of 0.58. The second best method, COS, had an average recall of 0.44 (Fig. 1A). The average precision of BioSim was 0.39, which was significantly higher than that of the other methods (P < 0.01, Wilcoxon signed-rank test). Likewise, BioSim had a median value of 2 for the FRR, surpassing the other methods (Table 1).

The overall performance of the methods varied for each benchmark category. It was lower for all methods when using the protein–protein interaction category and higher when using the sequence cluster category (Supplementary Fig. S3A and Table S1). The combined average recall for all methods was more than one-third lower in the protein–protein interactions category than in the sequence cluster category (the respective recalls were 0.29 and 0.75). The observed high performance when using the sequence clusters category is due to the tendency of the methods to rank similar sequences at the top. This can be explained, to some extent, by annotation transfer between homologous protein sequences, by gene annotations that are transferred to all encoded proteins and by domain annotations that are almost identical for similar sequences. Therefore, the tendency to rank similar sequences at the top reduces the performance of the methods when using benchmark categories different from sequence clusters because gold standard positives are displaced to lower ranks.

3.2 Including multiple annotation sources improves performance

The use of multiple annotation sources improved the performance of four of the five methods although they were not originally developed to handle multiple annotations (in contrast to BioSim). Much of this increase seems to be attributable to the availability of more annotation terms per protein. The number of terms annotated to each protein increased from a median of 7.5 GO terms to a median of 15.0 annotation terms when all annotation sources were included (Supplementary Figs S4A and S5A). The TO method, which counts the number of common terms, but does not account for term specificity, improved its average precision from 0.17 to 0.24 when all annotations were used.

Notably, the use of multiple annotation sources does not only increase the number of annotation terms per protein, but also improves the specificity of the annotations. While GO terms annotated to at most four proteins were available for 8096 proteins, this number doubled to 16 849 proteins in case of multiple annotation.

### Table 1. Performance comparison of functional similarity methods using multiple annotation sources versus using only GO annotations, over all 1243 validation groups

| Method | Multiple sources | Only GO |
|--------|-----------------|---------|
|        | Avg. precision | FRR     | Avg. precision | FRR     |
| BioSim | 0.39            | 2       | 0.22            | 7       |
| COS    | 0.28            | 3       | 0.22            | 7       |
| KC     | 0.21            | 5       | 0.20            | 5       |
| simGIC | 0.28            | 3       | 0.22            | 5       |
| TO     | 0.24            | 3       | 0.17            | 11      |

See Supplementary Table S1 for details on the performance of the methods in each of the four benchmark categories. avg. precision: average precision.
also reduce the impact of incorrect annotations found in biological databases (Schnoes et al., 2009).

3.3 BioSim scoring versus other methods

BioSim distinguished functional relationships of gold standard positives from those of randomly paired proteins better than the other methods. Gold standard positives consistently received a BioSim score close to 0, while random pairs obtained a score close to 1 (Fig. 2A). In particular, we plotted precision and recall averages from our benchmark results for every method at different score cut-offs (Fig. 2B). We also computed a score cut-off (SC50) that resulted in 50 false negatives on average. The obtained SC50 score cut-offs, along with the score range of each method from lowest to highest functional similarity, were as follows: BioSim: \( \leq 1.18 \times 10^{-9} \) (range [1:0]), TO: \( \geq 115 \) (range [0:∞]), KC: \( \geq 0.360 \) (range [0:1]), simGIC: \( \geq 0.096 \) (range [0:1]), and COS: \( \geq 0.101 \) (range [0:1]). For COS and simGIC, the second and third best methods, the SC50 score cut-offs were very close to zero, their non-similarity score; the recall at the respective SC50 cut-off had a median of 0.50 and a distribution covering the whole range (Fig. 2B). In other words, for both methods, the same SC50 cut-off resulted in a different recall. The KC and TO methods had a recall median below 0.5 for their respective SC50 score cut-offs. In comparison, the recall for BioSim at the SC50 score cut-off had the highest median (0.82) and the corresponding distribution concentrated around high values.

The limited consistency of the scores of COS, KC, simGIC and TO is probably caused by annotation bias toward better studied molecules (Rhee et al., 2008) as these methods appear to be best suited for unbiased data (Wang et al., 2010). In our data warehouse, a handful of proteins have over thousand annotations, while the majority has less than 10 annotations. A similar pattern can be observed when considering only GO annotations (Supplementary Figs S4 and S5). About 16% of all proteins are annotated only with less specific terms such as the UniProtKB keyword ‘Receptor’ or the GO term ‘protein binding’. The functional similarity of any two proteins sharing such terms is overestimated by the COS, KC and simGIC methods, which yield the highest score of 1. This misleading result is indistinguishable from a genuine functional similarity based on several shared annotation terms.

Furthermore, the same methods tend to underestimate the genuine similarity of any two proteins that are annotated with numerous terms and do not share a large proportion of their annotation terms. For example, the cellular tumor antigen TP53 (with 1642 annotation terms and do not share a large proportion of their annotation terms. Furthermore, useful annotations to derive functional similarities such as protein-protein interactions and disease associations are not part of GO. Moreover, the use of multiple annotation sources can
Novel search method for the discovery of functional relationships

Fig. 2. Comparison of functional similarity methods. (A) Histograms of the functional similarity scores that were obtained for 6907 pairs of gold standard positives and for 10 000 random pairs. (B) Precision (straight lines) and recalls (dashed lines) are averaged at different cut-offs. The vertical red lines highlight the SC50 score cut-offs that yield, on average, 50 false positives. The box plot to the left of the y-axis shows the distribution of recalls at this cut-off. BioSim scores are in logarithmic scale for better visualization. (C) Functional similarity and sequence similarity scores are compared based on 100 000 random pairs of proteins. Sequence similarity is measured as ln(bit score). Green lines depict the average functional similarity. Red lines illustrate the standard deviation. In each plot, the background contains a scatter plot where darker colors indicate a higher density of dots.

The TO method, which is simply the count of annotation terms shared by two proteins, avoids some of the described shortcomings by focusing only on the shared annotations. However, it cannot distinguish those annotations that occur by accident because it judges an event of two proteins sharing a rather unspecific, frequent annotation term (e.g. ‘protein binding’) as likely as an event of two proteins sharing a very specific, rare annotation term (e.g. ‘actin filament binding’).

3.4 Comparing functional similarity with sequence similarity

The correlation between the functional similarity of two proteins and their sequence similarity is often used to evaluate functional similarity methods (Lord et al., 2003; Pesquita et al., 2008). In our results, rank correlations for all methods were close to 0.1 when comparing BLAST bit scores and functional similarity scores for 100 000 random pairs of proteins. This low correlation is likely due to many protein pairs with almost no sequence similarity, but some functional similarity (Fig. 2C). To filter out protein pairs with low sequence similarity, we discarded all pairs having a ln(bit score) below 3.3. This threshold was chosen after observing that, for all methods, the averaged functional similarity scores increases above this value. In total, 631 (0.63%) of the random pairs had a ln(bit score) of at least 3.3. The rank correlations for these pairs were COS: 0.77, KC: 0.67, BioSim: 0.69, simGIC: 0.73, TO: 0.48.

Since BioSim showed a slightly lower correlation than COS and simGIC, we additionally analyzed some interesting cases manually. Supplementary Table S2 summarizes the manual inspection of annotations shared by the 15 pairs of proteins with the highest sequence similarity bit score. Seven protein pairs do not share specific annotation terms to infer a clear functional relationship. Accordingly, the low BioSim scores of those pairs are above the previously determined SC50 score cut-off of $1.18 \times 10^{-9}$, which indicates a weak functional similarity. In contrast, a true functional relationship between the remaining eight protein pairs is more evident due to several shared specific annotations terms. This agrees well with BioSim scores below or very close to the SC50 cut-off, which suggests a considerable certainty of a real functional similarity. However, in contrast to BioSim, the scores from the other methods do not allow a clear-cut distinction in those cases as explained in the preceding Section 3.3. For example, the 2nd and 15th rows in Supplementary Table S2 are cases of low functional similarity scores for COS, KC and simGIC in contrast to BioSim although the respective proteins share numerous annotations. This suggests that a meaningful comparison of scoring methods based on the correlation of functional similarity and sequence similarity is limited by the available annotation datasets and their overall characteristics and quality, which can also be affected by annotation bias and incompleteness. Since BioSim is particularly designed to be more sensitive to the number and specificity of annotation terms in contrast to the other methods, its overall performance depends more on the annotation datasets and the individual annotation terms.

3.5 Finding disease-associated genes

Genes associated with the same disease phenotype tend to be functionally related (Schlicker et al., 2010; Vidal et al., 2011).
Autosomal idiopathic short stature
Autosomal recessive
Antley–Bixler syndrome 1 FGFR1 Fibroblast growth factor receptor 1
Cardiofaciocutaneous syndrome 2 MAP2K1 Mitogen-activated protein kinase kinase 1
Folate-sensitive neural tube defects
Obesity 17 POMC Proopiomelanocortin
Autosomal recessive deafness-1A 1 GJB6 Gap junction protein, beta 6, 30 kD
Autosomal idopathic short stature
Hypogonadotropic hypogonadism 3 FGFR1 Fibroblast growth factor receptor 1
Non-insulin-dependent diabetes mellitus 25 PPARG Peroxisome proliferator-activated receptor gamma
Susceptibility to atypical hemolytic uremic syndrome-1 2 CFI Complement factor I
Non-insulin-dependent diabetes mellitus 25 SLC2A4 Solute carrier family 2 (facilitated glucose transporter), member 4

The table lists 12 new disease gene associations found between ranks 1 and 6. The table column ‘# genes’ gives the number of known genes associated with the disease phenotype before January 1, 2009. The column ‘New gene’ contains the symbol of the gene that was added to the phenotype between January and October 2009 and correctly identified by BioSim. The columns ‘Rank’ and ‘GO rank’ give the position of the new gene in the ranking list if all annotations were used or only GO, respectively. The column ‘Shared annotations’ contains a summary of the most specific annotation terms shared by the known genes and the new gene. The detailed list of shared annotations can be found in Supplementary Tables S3–S26. Gene symbols and descriptions correspond to the official nomenclature from HGNC (Seal et al., 2011). Indirect PPI refer to all direct interaction partners of the same protein.

Using BioSim, we ranked genes based on their functional similarity to genes known to be associated with a particular OMIM disease phenotype (Amberger et al., 2009). To this end, for each gene not associated with a disease phenotype, we averaged the computed scores of its functional similarity to the previously known disease genes. The functional similarity scores were computed using a snapshot of our data warehouse that contained only gene annotations before January 1, 2009. We then compared our results with an updated version of OMIM from October 31, 2009. This update contained 54 new gene associations for 46 diseases. In our results, 11 of the new genes were found at the top four ranks and 12 others between ranks 6 and 54 (Table 2 and Supplementary Tables S3–S26). The median rank of the new genes was 9.5. This is a drastic improvement due to the use of multiple annotation sources in contrast to the ranks obtained when using only GO annotations with a resultant median of 133.5.

Figure 3 highlights two disease phenotypes: obesity, which had 17 associated genes known before January 2009, and familial glioma of brain, which had seven associated genes. The new gene POMC, which was added to the obesity phenotype in the updated version of OMIM, was found on the third rank. Annotations shared by POMC and the other known disease genes included protein–protein interactions (with AGRP, ENPP1, GHR, MC3R and MC4R) and the annotation term ‘obesity’ from UniProtKB keywords, which covers POMC and 10 other obesity genes (Supplementary Table S9). The genes ranked first and second, LEP (leptin) and INS (insulin), are also related to obesity (Spiegelman and Flier, 2001) even if they are not among the genes of the specific obesity phenotype in OMIM.

BRCA2, the new gene included into the updated version of OMIM for the glioma of brain phenotype, achieved the first rank of genes functionally related to the disease. BRCA2 showed strong BioSim functional similarity to five of the seven previously known genes for glioma of brain. Some of the annotations shared by BRCA2 and the five disease genes are protein–protein interactions (with ERBB2, MSH2 and PTEN), the joint disease association of BRCA2 and DMRT1 to medulloblastoma as well as of BRCA2 and PTEN to prostate cancer in OMIM and a number of GO and pathway annotations (Supplementary Table S3).

4 CONCLUSION
We presented the novel method BioSim to compute and search for functional similarities of genes and proteins based on diverse annotations such as protein interactions, domain architectures,
The research was conducted in the context of the DFG-funded Cluster of Excellence for Multimodal Computing and Interaction. German Research Foundation (DFG) contract number KFO 129/1-2.

**ACKNOWLEDGEMENTS**

We are grateful to Adrian Alexa for his advice on R optimization and to Andreas Schlicker for sharing his knowledge on similarity methods.

**Funding**

German National Genome Research Network (NGFN); German Research Foundation (DFG) contract number KFO 129-1-2. The research was conducted in the context of the DFG-funded Cluster of Excellence for Multimodal Computing and Interaction.

**Conflict of Interest:** none declared.

**REFERENCES**

Aerts,S. et al. (2006) Gene prioritization through genomic data fusion. Nat. Biotechnol., 24, 531–544.

Altschul,S.F. et al. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410.

Amberger,J. et al. (2009) McKusick’s Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res., 37, D703–D709.

Bakal,O.. (2007) Nature Milestones in DNA technologies, Milestone 15: BLAST-off for genomes. Nat. Rev. Genet., 8, S14–S15.

Barth,A. (2000) The ENZYME database in 2000. Nucleic Acids Res., 28, 304–305.

Bennetzen,J.L. et al. (2010) InteFGO: a new vector-based semantic similarity measure including annotation origin. BMC Bioinformatics, 11, S88.

Berman,H. et al. (2003) Announcing the worldwide Protein Data Bank. Nat. Struct. Biol., 10, 980.

Blake,J.A. et al. (2011) The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Res., 39, D842–D849.

Buckley,C. and Voorhees,E.M. (2000) Evaluating evaluation measure stability. In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000). ACM, New York, NY, USA, pp. 33–40.

Butte,A.L. et al. (2008) (ICAM) is a novel Rab2 effector regulating ER-Golgi trafficking in insulinoma cells. Eur. J. Cell Biol., 87, 197–209.

Camon,E. et al. (2004) The Gene Ontology Annotation (GOA) Database – an integrated resource of GO annotations to the UniProt Knowledgebase. In Silico Biol., 4, 5–6.

Chabalier,J. et al. (2007) A transversal approach to predict gene product networks from ontology-based similarity. BMC Bioinformatics, 8, 235.

Chan,J.N.Y. et al. (2010) Recent advances and method development for drug target identification. Trends Pharmacol. Sci., 31, 82–88.

Chen-Aryamontri,A. et al. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res., 35, D572–D574.

Chen,F. et al. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res., 34, D663-D668.

Consortium,U. (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res., 38, D142–D148.

del Pozo,A. et al. (2008) Defining functional distances over Gene Ontology. BMC Bioinformatics, 9, 50.

Finn,R.D. et al. (2008) The Pfam protein families database. Nucleic Acids Res., 36, D212–D218.

Flueck,P. et al. (2008) Ensembl 2008. Nucleic Acids Res., 36, D707–D714.

Freundl,J. (2006) Automated protein function prediction-the genomic challenge. Brief. Bioinform., 7, 225–242.

Gelbukh,A. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput. Chem., 20, 25–33.

Huang,D.W. et al. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large lists. Genome Biol., 8, R183.

Hunter,S. et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res., 37, D211–D215.

Huttenhower,C. et al. (2009) Exploring the human genome with functional maps. Genome Res., 19, 1103–1106.

Jensen,J.L. et al. (2009) STRING 8 – a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res., 37, D412–D416.

Kanehisa,M. et al. (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res., 36, D480–D484.

Kerrien,S. et al. (2007) IntAct – open source resource for molecular interaction data. Nucleic Acids Res., 35, D615–D619.

Lenz,G. and Shakhnovich,E.B. (2007) Defining functional distance using manifold embeddings of Gene Ontology annotations. Proc. Natl Acad. Sci. USA, 104, 11334–11339.

Liu,D. (1998) An information-theoretic definition of similarity. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998). ACM, New York, NY, USA.

Liu,J.P. et al. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 19, 1275–1283.

Matthews,L. et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res., 37, D619–D622.

McMahon,H.T. et al. (1995) Complexins: cytosolic proteins that regulate SNAP receptor function. Cell, 83, 111–119.

Minty,M. and Pavlidis,P. (2008) Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics, 9, 327.

Pavlidis,P. et al. (2008) Metis for GO-based protein semantic similarity: a systematic evaluation. BMC Bioinformatics, 9, S48.

Popescu,M. et al. (2006) Fuzzy measures on the Gene Ontology for gene product similarity. IEEE/ACM Trans. Comput. Biol. Bioinform., 3, 263–274.
Schnoes, A.M. et al. (2009) Human Protein Reference Database – 2009 update. Nucleic Acids Res., 37, D767–D772.
Ramirez,F. et al. (2007) Computational analysis of human protein interaction networks. Proteomics, 7, 2541–2552.
Reeves,G.A. et al. (2008) The Protein Feature Ontology: a tool for the unification of protein feature annotations. Bioinformatics, 24, 2767–2772.
Rosci,P. (1995) Using information content to evaluate semantic similarity in a Taxonomy. In Proceedings of the 4th International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada, Morgan Kaufmann, San Francisco, CA, USA, pp. 448–453.
Rosci,P. (1999) Semantic similarity in a Taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res., 11, 95–130.
Rhier,S.Y. et al. (2008) Use and misuse of the gene ontology annotations. Nat. Rev. Genet., 9, 309–315.
Rhodes,D.R. et al. (2005) Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol., 23, 951–959.
Romero,P. et al. (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol., 6, R2.
Rupp,A. et al. (2008) CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res., 36, D646–D650.
Salton,G. et al. (1975) A vector space model for automatic indexing. Commun. ACM, 18, 613–620.
Salwinski,L. et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res., 32, D449–D451.
Schlicker,A. et al. (2006) A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics, 7, 302.
Schlicker,A. et al. (2010) Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics, 26, i561–i567.
Schnee, A.M. et al. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput. Biol., 5, e1000605.
Seal,R.L. et al. (2011) genenames.org: the HGNC resources in 2011. Nucleic Acids Res., 39, D514–D519.
Sevilla,J.L. et al. (2005) Correlation between gene expression and GO semantic similarity. IEEE/ACM Trans. Comput. Biol. Bioinform., 2, 336–338.
Sing,T. et al. (2005) ROC: visualizing classifier performance in R. Bioinformatics, 21, 3940–3941.
Spence,N. et al. (2004) A memetic clustering algorithm for the functional partition of genes based on the Gene Ontology. In Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004), La Jolla, CA, USA, IEEE Press, San Diego, CA, USA, pp. 252–259.
Spiegelman,M. and Flier,J.S. (2001) Obesity and the regulation of energy balance. Cell, 104, 531–547.
Suek,B.E. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 23, 1282–1288.
Su,A.I. et al. (2002) Large-scale analysis of the human and mouse transcriptomes. Proc. Natl Acad. Sci. USA, 99, 4465–4470.
Vasilev,L.T. et al. (2004) In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. Science, 303, 844–848.
Velankar,S. et al. (2005) E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res., 33, D262–D265.
Vidal,M. et al. (2001) Intrachain networks and human disease. Cell, 144, 986–998.
Wang,J. et al. (2010) Revealing and avoiding bias in semantic similarity scores for protein pairs. BMC Bioinformatics, 11, 290.
Wang,F. and Marcotte,E.M. (2010) It’s the machine that matters: Predicting gene function and phenotype from protein networks. J. Proteomics, 73, 2277–2289.
Watts-Fairley,D. et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res., 38, W214–W220.
Willott,P. et al. (1998) Chemical Similarity Searching. J. Chem. Informat. Comput. Sci., 38, 983–996.