Methods in Molecular Biology (2016) 1375: 117–121
DOI 10.1007/7651_2015_249

Published online: 14 May 2015

Ontology-Based Analysis of Microarray Data

Agapito Giuseppe and Marianna Milano

Abstract

The importance of semantic-based methods and algorithms for the analysis and management of biological data is growing for two main reasons. From a biological side, knowledge contained in ontologies is more and more accurate and complete, from a computational side, recent algorithms are using in a valuable way such knowledge. Here we focus on semantic-based management and analysis of protein interaction networks referring to all the approaches of analysis of protein–protein interaction data that uses knowledge encoded into biological ontologies.

Semantic approaches for studying high-throughput data have been largely used in the past to mine genomic and expression data. Recently, the emergence of network approaches for investigating molecular machineries has stimulated in a parallel way the introduction of semantic-based techniques for analysis and management of network data. The application of these computational approaches to the study of microarray data can broad the application scenario of them and simultaneously can help the understanding of disease development and progress.

Keywords: Data mining, Expression patterns, Bi-clustering, Microarray

1 Introduction

The accumulation of data about proteins, genes, and small molecules on a large scale caused the possibility to look at molecular machineries on a system level scale. After the rise of the systems biology, more recently the network biology (1), i.e., the discipline that bring together molecular biology and network theory, has gained a big interest.

In this scenario, data about genes constitute the fundamental building blocks (2) used to grow models and theories.

Let us consider for instance interactions among proteins, named protein–protein interactions (PPI). Proteins play their role usually by interacting with them or other macromolecules. An interaction usually involves a contact with surfaces of two or more proteins.

Due to the introduction of high-throughput techniques, many experimental datasets have been produced causing the introduction of computer science methods to manage, store, and analyze PPI data (3). The whole set of protein interactions of a single species are also referred to as Protein to Protein Interaction Network (PIN). PINs have been easily modeled by using undirected graphs (4) where nodes are associated with proteins and edges represent interactions among proteins.

PPI data have been collected in many public databases.

Usually, PPI databases contain raw data, e.g., the identifiers of the interacting proteins, and some annotation related to the reliability of the stored data.

The accumulation of raw experimental data about genes and proteins have been accompanied by the accumulation of functional information, i.e., knowledge about function. The assembly, organization, and analysis of this data have given a considerable impulse to research (5). Usually, biological knowledge is encoded by using annotation terms, i.e., terms describing for instance function or localization of genes and proteins. Such annotations are often organized into ontologies, which offer a formal framework to organize in a formal way biological knowledge.

For instance, Gene Ontology (GO) (6) provides a set of annotations (namely GO Terms) of biological aspects, structured into three main taxonomies: Molecular function (MF), Biological Process (BP), and Cellular Component (CC). Annotations are often stored in publicly available databases, for instance, a main resource for GO annotations is the Gene Ontology Annotation (GOA) database (7). The availability of well-formalized functional data enabled the development of algorithms and methods to analyze proteins and genes from a semantic perspective.

Historically, first approaches were referred to as functional enrichment algorithms. They have been developed to determine the statistical significance of the presence (or the absence) of a GO term in a set of gene products or proteins (8). Despite the existence of more than 60 freely available tools, the functional analysis of large list is still a challenge. Classical algorithms referred to Gene Enrichment Algorithms (GEA) or Gene Set Enrichment Algorithms (GSEA), do not cope with the topological information contained in protein or gene interaction network. More recently, network enrichment analysis (NEA) approaches that extends the classical approaches to network links between genes in the experimental set and those in the functional categories (9).

More recently, a set of algorithms, referred to as Semantic Similarity Measures (SSMs), have been developed to compare in a quantitative way set of terms belonging to the same ontology. SSMs take in input two or more ontology terms and produce as output a value representing their similarity.

This enabled the possibility to use such formal instruments for the comparison and analysis of proteins and genes (10, 11). Many works have focused on: (1) the definition of ad-hoc semantic measures tailored to the biological scenario (12); (2) the introduction of algorithms for the functional analysis of interactomics data (13);

and (3) finally the building of semantic similarity networks (SSN), i.e., edge-weighted graph whose nodes are genes or proteins, and edges represent semantic similarities among them (14).

2 Semantic Similarity Measures

While sequence or structure-based similarity of genes and proteins has been largely investigated in the past, the similarity based on functions presents a more complex scenario. In fact, while primary and tertiary structures can be compared in terms of number of shared amino acids or in terms of spatial conformation. The comparison of the functions needs the introduction of a comparison metrics between terms that are often expressed in natural language.

The adoption of ontologies for managing annotations provides a means to compare entities on aspects that would otherwise not be comparable. For instance, if two gene products are annotated within the same schema, we can compare them by comparing the terms with which they are annotated (15, 16).

The annotations of biological concepts are currently organized in simple taxonomies or more complex ontologies, such as Gene Ontology or Open Biomedical Ontologies (OBO), (17). The use of ontologies enables the comparison of annotations in terms of analysis of the ontology schema. Thus, the problem to define the semantic similarity of two terms can be solved in terms of analysis of the underlying ontology. While the semantic similarity among two biomedical or biological concepts is not a trivial problem, the semantic similarity among terms that come from a common schema, e.g., a taxonomy has been largely investigated and can be solved in an efficient way. In the same way, if two biological concepts, e.g., proteins are annotated with terms organized by using an ontology, the problem of the determination of their semantic similarity can be solved in terms of semantic similarity of the annotating terms.

Several approaches are available to quantify the semantic similarity between terms or annotated entities in an ontology represented as a directed acyclic graph (DAG) such as GO.

We here presents a brief categorization on the basis of according to the strategy used for the calculation: (1) Term Information Content (IC), (2) Term Depth, (3) based on a common ancestor, (4) based on all common ancestors, (5) Path Length, and (6) Vector Space Models (VSM). Measures based on Term Depth and IC evaluate terms similarity on the basis of the specificity of the terms. Measures based on a common ancestor first select a common ancestor of two terms according to its properties, and then evaluates the semantic similarity on the basis of the distance among the terms and their common ancestor and the properties of the common ancestor. Techniques based on Path Length correlate measures of the length of the path connecting the two terms. VSM-based measures initially represent the set of the annotations of proteins as vectors. Then, the similarity is evaluated by considering the distance among vectors that are defined using topological considerations.

Proteins and genes are annotated with a set of GO terms, so to assess the functional similarity between gene products it is necessary to compare sets of terms rather than single terms. All the proposed approaches are based on the comparison of terms and on the combination of the results, i.e., the pairwise similarity of annotations calculated using an existing measure. The simplest way to measure the semantic similarity between two gene products is to calculate the pairwise semantic similarity among the terms that annotate the gene products and successively to combine such pairwise similarity by using some formulas such as the average, the maximum, or the sum. Other approaches are based on the representation of two gene products as the induced subgraph of annotation or as a point in a vector space induced by annotations (18, 19).

Semantic similarity measures are affected by three main problems (20):

Annotation length. The number of annotations per protein (i.e., the GO Terms associated with each protein) is highly variable within the same GO taxonomy and over different species. Consequently, the resulting similarity score is affected by this variability. Consequently, comparing proteins with few annotations is more likely to return low similarity scores, even if the proteins are related.

Evidence codes. The task of associating with proteins the GO Terms that describe their functions and properties, called annotation, is performed with different methods. Without entering into details, they range from experimentally verified to electronIcally infErred Annotations (IEA). SSMs usually do not weight annotations on the basis of their ECs, and one has to choose between including potentially unreliable annotations to increase the number of annotations at the expenses of the quality or ignoring them but drastically reducing the number of annotations considered.

Shallow annotations. Several proteins are annotated with generic GO terms. These annotations do not identify the specific role or function of the protein, but only suggest the area in which the proteins operate.

References

1. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113
2. Cannataro M, Guzzi PH, Veltri P (2010) Protein-to-protein interactions: technologies, databases, and algorithms. ACM Comput Surv. doi:10.1145/1824795.1824796
3. Ciriello G et al (2012) AlignNemo: a local network alignment method to integrate homology and topology. PLoS One. doi:10.1371/journal.pone.0038107
4. West DB (2000) Introduction to graph theory, 2nd edn. Prentice Hall, New York
5. Blake JA, Bult CJ (2006) Beyond the data deluge: data integration and bio-ontologies. J Biomed Informat 39(3):314–320
6. Harris MA et al (2004) The gene ontology (go) database and informatics resource. Nucleic Acids Res 32:258–261
7. Barrell D et al (2009) The GOA database in 2009-an integrated gene ontology annotation resource. Nucleic acids Research. doi:10.1093/nar/gkn803
8. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1–13
9. Alexeyenko A et al (2012) Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinformatics. doi:10.1186/1471-2105-13-226
10. Guzzi PH et al (2012) Semantic similarity analysis of protein data: assessment with biological features and issues. Brief Bioinform 13(5):569–585
11. Smoot ME et al (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27(3):431–432
12. Pesquita C et al (2009) Semantic similarity in biomedical ontologies. PLoS Comput Biol. doi:10.1371/journal.pcbi.1000443
13. Dai X et al (2014) A comprehensive semantic similarity measurement for predicting the function of gene products. J Bionanosci 8 (4):287–292
14. Agapito G, Guzzi PH, Cannataro M (2013) Visualization of protein interaction networks: problems and solutions. BMC Bioinformatics. doi:10.1186/1471-2105-14-S1-S1
15. Guzzi PH, Cannataro M (2012) Cyto-sevis: semantic similarity-based visualisation of protein interaction networks. EMB-Net J doi: http://dx.doi.org/10.14806/ej.18.A.397
16. Cannataro M et al (2007) Using ontologies for preprocessing and mining spectra data on the grid. Future Generat Comput Syst 23 (1):55–60
17. Smith B et al (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25 (11):1251–1255
18. Popescu M, Keller JM, Mitchell JA (2006) Fuzzy measures on the gene ontology for gene product similarity. IEEE/ACM Trans Comput Biol Bioinformatics 3 (3):263–274
19. Yu G, Li F et al (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 26(7):976–978
20. Guzzi PH, Mina M (2012) Towards the assessment of semantic similarity analysis of protein data: main approaches and issues. ACM SIGBioinformatics Rec 2(3):17–18