Bioinformatic tools for proteomic data analysis: an overview

REVIEW

✍ Arielis Rodríguez-Ulloa, Rolando Rodríguez

Physical Chemistry Division, Center for Genetic Engineering and Biotechnology, Havana, Cuba
Ave 31 e/ 158 and 190, Playa, PO Box 6162, Havana 10 600
E-mail: arielis.rodriguez@cigb.edu.cu

ABSTRACT

Over the years the identification of one gene or protein has been substituted by the determination, in a single experiment, of all the genes or proteins expressed in a given cellular state. Consequently, the amount of available data has increased considerably. In this scenario, bioinformatics has become the bridge between experimental data and computational tasks for managing, mining and retrieving information. The current paper provides an overview on the bioinformatic databases and software used to analyze the biological meaning of proteomic data, thus describing the main functions and limitations of these tools. The important challenges of data analysis are mainly related to the integration of biological information from dissimilar sources. In this field, the improvement in databases and developments in software in the future would contribute to the potential opportunities given by proteomics.

Keywords: Proteomic data analysis, systems biology, bioinformatics, biological databases

Biotecnología Aplicada 2008;25:312-319

RESUMEN

Herramientas informáticas para el análisis de datos en proteómica: una perspectiva. La posibilidad de identificar todos los genes o proteínas expresados en un determinado estado celular ha incrementado considerablemente la cantidad de datos disponibles. Por tanto, la bioinformática se ha convertido en un puente que media entre los datos experimentales y las tareas computacionales de gestión, minería y recuperación de información. En este artículo se mencionan algunas bases de datos y programas bioinformáticos que se utilizan en la interpretación biológica de los datos obtenidos en experimentos de proteómica. Acerca de estas herramientas se describen sus principales funcionalidades y limitaciones. Los retos actuales del análisis de datos están relacionados principalmente con la integración de información biológica a partir de diferentes fuentes. En este sentido, deben desarrollarse nuevas bases de datos y programas, que en su conjunto contribuyan a lograr las oportunidades potenciales que ofrece la proteómica.

Palabras clave: Análisis de datos proteómicos, biología de sistemas, bioinformática, bases de datos biológicas

*I*ntroduction

Scientific research has changed in recent years mainly due to the completion of numerous genomes in addition to the development and application of high-throughput technologies including gene expression microarrays and mass spectrometry. The identification of all expressed gene transcripts, proteins, or metabolites, termed the transcriptome, proteome and metabolome, respectively [1], has imposed new challenges because of the large amount of data that needs a comprehensive analysis, but at the same time it has also brought about new and increasing opportunities to enhance the knowledge on biological systems as a whole.

Nevertheless, the identification and quantification of proteins or gene transcripts separately is not enough to fully understand functional changes in a given cellular state. For this, research at different levels and also the integration of different types of information is required. In this sense a new discipline, known as systems biology has evolved. It combines experimental approaches with computational biology to help understand the biological phenomena on a global scale [2].

Proteins are the main catalysts, structural elements, signaling messengers and molecular machines of a cell, which is a strong argument to support the advantages and importance of directly analyzing proteins. Proteomics is defined as the large scale identification and functional characterization of all expressed proteins in a given cell (in a given state), including all protein isoforms and modifications, protein interaction networks, protein structure determination and high-order complexes of proteins [3]. An important progress in proteomics has been achieved by the introduction of powerful new technologies and high throughput experiments. Several reviews address the advancing technology available for proteomic studies [3-6].

Specifically, an analysis starting with proteomic data and taking into account the integration needed to understand the system is what we name a post-proteomic study. Based on the principles of systems biology, this type of study is an important complement to achieve the goals of any proteomic research. During a post-proteomic analysis, bioinformatic tools such as databases and software are indispensable (Figure 1). It is beyond the scope of this review to describe all the databases or software available for data analysis or to give a methodology for interpreting proteomic data; first of all, because biological data are distributed in a large number of databases that can not be described in a single paper. Also a post-proteomic methodology must be based on an investigation hypothesis and the questions that must be answered in each study. In any case, to discover new drug targets, predict the potential toxicity of drugs, propose disease biomarkers or in general to understand any biological process, some basic points must be followed. This paper, therefore, mentions those aspects which are the core of a post-proteomic study and gives an overview of the tools that can be used to meet these general objectives,

1. Oliver S. Proteomics: Guilt-by-association goes global. Nature 2000;403: 601-3.

2. Kitano H. Computational systems biology. Nature 2002;420:206-10.

3. Tyers M, Mann M. From genomics to proteomics. Nature 2003;422:193-7.

4. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198-207.

5. Domon B, Aebersold R. Mass spectrometry and protein analysis. Science 2006;312:212-7.

6. Ong SE and Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 2005;1:252-62.

✍ Corresponding author

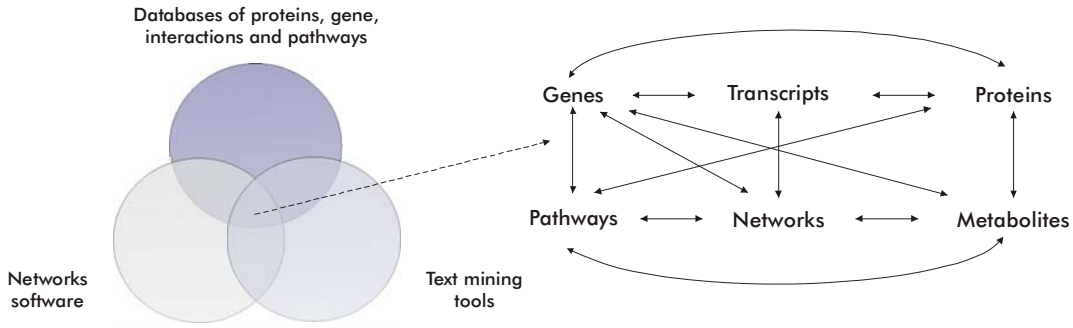

Figure 1. During a post-proteomic analysis, the combination of different data sources and the application of other bioinformatic tools enable the integration among different biological levels.

although these tools are not limited to proteomic studies and can clearly be used in any kind of data analysis. Finally the basic limitations and future developments in the field are pointed out.

Post-proteomic tools: databases and bioinformatic software

Beginning with essential databases

The final goal of a post-proteomic study determines which databases should be used. For example, in cancer projects, to identify proteins that might be biomarkers of anti-cancer drugs a direct correlation between the differential expression profiles of the potential biomarkers and drug action is essential. To know if the proteins have been previously identified in cancer samples, an initial resource could be the Oncomine database which contains a collection of cancer gene expression profiles [7]. However these results must be analyzed carefully because expression changes at transcript and protein level may not always be correlated [8]. To identify potential cancer targets besides using Oncomine, Cosmic database also has valuable information. This resource stores data on somatic mutations in different cancer types [9]. Also toxic effects of new drugs could be predicted if proteomic results are related to adverse reactions of known drugs. In this sense, even if the project is not related with cancer, the molecular bases of drug action could be studied using the SuperTarget database which contains drug-related information such as targets, pathways affected and adverse effects [10]. In proteomic studies of diseases caused by microorganisms, such as those intended to identify targets or understand the molecular basis of the infection, the Comprehensive Microbial Resource (CMR) could be used. This resource has information on complete prokaryotic genomes and enables their comparison [11]. Thus CMR makes it possible to compare different microorganism genomes to find the genes coding homologous proteins. Those proteins must be the targets of an anti-microbial drug to ensure a wide spectrum of action. Also putative targets must be essential in the survival of the pathogen, and highly divergent from human genes.

A final example is the HIV-1 - Human Protein Interaction database, which is helpful to understand the processes of HIV-1 (human immunodeficiency virus type 1) replication and pathogenesis at a proteomic level [12]. Other protein interaction databases will be dealt with later.

In any post-proteomic project, however, it is important to know basic, but essential, information on the experimentally identified proteins. For example: Which are the coding genes? How are they regulated? How do post-translational modifications affect protein function? Is the protein related to any disease? The answers to these questions are in databases such as Uniprot, EntrezGene and OMIM (Table 1, Figure 2). This step is the core of the analysis and could become a bottleneck if the dataset is from a large-scale experiment. At this point a bioinformatic procedure, for automatic searches in these databases and for saving the results for their subsequent analysis, is an advantage for finding relevant information.

Proteins can be grouped according to their molecular function or biological process. A convenient classification system is obtained from the Gene Ontology (GO) database, which has a defined and controlled vocabulary to describe the roles of genes and gene products in any organism (Table 1, Figure 2). Three independent ontologies are accessible: biological process, molecular function and cellular component [13]. Biological process refers to a biological objective to which the gene or gene product contributes. It is accomplished via one or more ordered assemblies of molecular functions. Thus, the molecular function is

7. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, *et al*. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 2007;9:166-80.

8. Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 1997;18: 533-7.

9. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, *et al*. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 2004;91:355-8.

10. Gunther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, *et al*. SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res 2008;36:D919-D922.

11. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The Comprehensive Microbial Resource. Nucleic Acids Res 2001;29:123-5.

12. http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/2008.

13. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25-9.

Table1. Databases of proteins and genes

DB name	Description	Reference URL	A_V	C_V
UniProt Knowledgebase	Curated protein sequence database with a minimal level of redundancy and a high level of integration with other databases	[14] http://br.expasy.org/sprot/	Yes	No
Entrez Gene	NCBI database with gene specific information and links to citations and to external database	[15] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene	Yes	No
OMIM Online Mendelian Inheritance in Man	Catalog of human genes and genetic disorders that focus primarily on inherited or heritable genetic diseases. It contains textual information, references and links to MEDLINE and sequence records in the Entrez system of NCBI	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM	Yes	No
Gene Ontology	Provides a controlled vocabulary to describe gene and gene product attributes in any organism	[13] http://www.geneontology.org/	Yes	No

NCBI: National Center for Biotechnology Information, DB: Databases, A_V: Academic Version, C_V: Commercial Version

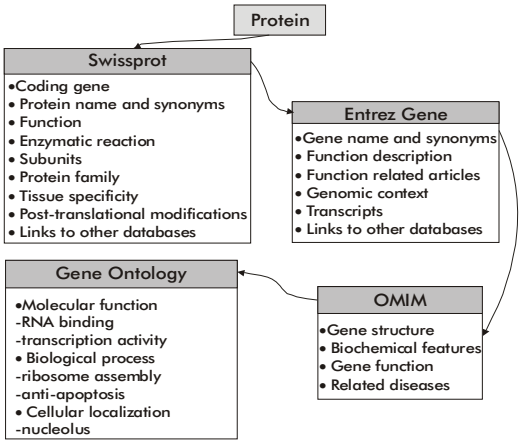

Figure 2. Databases that can be consulted not only during post-proteomic analysis but also in any kind of data analysis project. Shown in each case is some of the information that can be retrieved, specifically each category of Gene Ontology that is exemplified with associated terms. These databases have links between them and although any database could be the starting point in the analysis, a possible workflow is depicted with arrows.

defined as the biochemical activity (including specific binding to ligands or structures) of a gene product. Finally, the cellular component refers to the location in the cell where a gene product is active. To facilitate the identification of GO terms associated with the experimental data, tools such as Golem and GOtreePlus are available [16, 17]. Golem is independent of the operating system while GOtreePlus is only available for Windows users.

Protein interaction databases

Protein-protein interactions are necessary for almost all cellular functions [18]. To determine protein interactions, experimental methods have been used such as the yeast 2 hybrid screen test [19], co-occurrence under specific conditions of gene expression [20], co-immunoprecipitation of protein complexes [21], and tandem affinity purification [22, 23]. The interactions supported by experimental evidence are included directly in the databases or incorporated from the literature into repositories through the curation of biologists.

Repositories of protein interactions from different organisms are available in databases such as DIP, BIND, Intact and MINT (Table 2). Furthermore, the MIPS database can be used for mammalian protein interactions (Table 2). There are also specific databases for the analysis of human protein interactions, for example HPRD and HomoMINT (Table 2).

HPRD, besides protein interactions, also provides information on: protein isoforms, domain architecture, protein functions, post-translational modifications, subcellular localization and tissue expression [24]. It also includes a tool called PhosphoMotif Finder, which searches the presence of any of the more than 320 phosphorylation motifs, taken from the literature, in proteins sequences [25]. On the other hand, HomoMINT contains interactions obtained by transferring the experimental interaction annotation from the proteome of model organisms, such as *Saccharomyces cerevisiae* and *Drosophila melanogaster*, to the corresponding orthologous human proteins [26]. Since it is not clear what percentage of the protein-protein interactions is conserved through evolution [27], HomoMINT should be considered carefully only as a resource of hypothetical interactions. The STRING database also contains annotations of protein-protein interactions inferred by data obtained in other organisms in which there are homologous protein pairs that interact [28] (Table 2). Thus, HomoMINT and STRING could be considered general exploratory resources. They are best used for a quick initial overview of the functional partners of a query protein, especially for non-characterized proteins.

The protein interaction databases cited in this paper have been previously compared except for STRING and HomoMINT. Although they overlap well at the protein level, the protein-protein interaction data do not overlap as much [29]. For example protein overlapping between BIND (3887 proteins) and IntAct (4614 proteins) is 1969 but the overlap at the protein-protein interaction level is of only 1167 (the figures refer to the date of Mathivanan's study, October 2006) [29]. Therefore, the information from these databases complement each other and together they can increase and improve our knowledge on interactome networks.

Efforts have been made, in recent years, to integrate in one place the data contained in different databases. Two recent approaches are APID (Agile Protein Interaction Data Analyzer) and MIMI (Michigan Molecular Interactions). APID integrates data from five main source databases: BIND, DIP, HPRD, IntAct and MINT [30], while MIMI includes the rest of these source databases except for MINT [31]. Moreover MIMI has other interaction sources such as the General Repository for Interaction Datasets (BioGrid) [32]. Data from both databases can be retrieved, visualized and handled with Cytoscape plugins. However the APID plugin, named APID2NET, has additional functions compared to the MIMI plugin. In a network constructed with APID2NET the proteins can be highlighted according to their related families or GO terms (Figure 3).

The distribution of interaction partners differs across different databases [29]. A thoroughly studied human protein such as caspase 3 (CASP3) has 126 protein interaction partners annotated in HPRD, while MINT, DIP and IntAct just contain 11, 0 and 4 interactions, respectively. Often only one isoform is annotated in databases as interacting although there is no evidence that the interaction is specific to that isoform [29], as expressed by Mathivanan *et al*., most of the interaction databases are still at an early stage of curation and annotation of published protein-protein interactions. Therefore, reference articles are needed to validate the analyzed data, which is essential during a post-proteomic analysis, to be able to understand the functional relevance of the interactions. Moreover, published papers are a source of up-to-date information which is frequently not found in databases. But papers in this field grow exponentially. In fact, a concern of the scientific community is how to overcome the overwhelming amount of available information.

Searching information with text mining tools

Text mining algorithms have been used to find co-occurrence, and therefore the association of gene or protein names in the same text. For example, Natural

14. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supple-ment TrEMBL in 2000. Nucl Acids Res 2000; 28:45-8.

15. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucl Acids Res 2005;33:D54-8.

16. Sealfon RS, Hibbs MA, Huttenhower C, Myers CL, Troyanskaya OG. GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. BMC Bioinformatics 2006;7:443.

17. Lee B, Brown K, Hathout Y, Seo J. GOTreePlus: an interactive gene ontology browser. Bioinformatics 2008;24:1026-8.

18. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature 2000;405:823-6.

19. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Si USA 2001; 98:4569-74.

20. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, *et al*. Functional Discovery via a Compendium of Expression Profiles. Cell 2000; 102:109-26.

21. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415:141-7.

22. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B. A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 1999;17:1030-2.

23. Forler D, Kocher T, Rode M, Gentzel M, Izaurralde E, Wilm M. An efficient protein complex purification method for functional proteomics in higher eukaryotes. Nat Biotechnol 2003;21:89-92.

24. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, *et al*. Human protein reference database-2006 update. Nucl Acids Res 2006;34:D411-D414.

25. http://www.hprd.org/PhosphoMotif_finder.

26. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. BMC Bioinformatics 2005;6 Suppl 4:S21.

27. Cesareni G, Ceol A, Gavrila C, Palazzi LM, Persico M, Schneider MV. Comparative interactomics. FEBS Lett 2005;579: 1828-33.

28. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, *et al*. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucl Acids Res 2005;33:D433-7.

29. Mathivanan S, Periaswamy B, Gandhi TK, Kandasamy K, Suresh S, Mohmood R, *et al*. An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics 2006;7 Suppl 5:S19.

30. Prieto C, De Las RJ. APID: Agile Protein Interaction DataAnalyzer. Nucleic Acids Res 2006;34:W298-302.

Table 2. Protein interactions databases. In each case, besides the description, we point out the additional information that can be found in these databases and if an interaction network representation is available. (DB: Databases, Int_Netw: Interaction Network, A_V: Academic Version, C_V: Commercial Version)

DB name	Description	Int_Netw representation	Reference URL	A_V	C_V
DIP Database of Interacting Proteins	Experimentally determined protein–protein interactions. Gives the position of an interaction in a pathway and specific post-translational modifications	Available	[33] http://dip.doe-mbi.ucla.edu/	Yes	Yes
BIND Biomolecular Interaction Database	Molecular interactions (protein-protein, protein-nucleic acid, protein-small molecule) derived from the literature and experimental datasets. Also has information on protein complexes and pathways	The pathways can be exported and visualized in Cytoscape	[34] http://www.bind.ca	Yes	Yes
Intact	Interactions are derived from literature curation or direct user submissions. Interacting domains and experimental method used are also annotated	Available through the applications ProViz and Hierarch View	[35] http://www.ebi.ac.uk/intact/site/index.jsf	Yes	No
MIPS Munich Information Center for Protein Sequences	Mammalian protein-protein interactions extracted manually from scientific literature. Data from mass spectrometry and yeast-two-hybrid studies are not included. Gives the probable binding regions of interacting partners, the functional role of the interaction and access to mammalian protein complexes database (MPCDB)	-	[36] http://mips.gsf.de/	Yes	No
HPRD Human Protein Reference Database	Human protein–protein interactions annotated manually from the literature or by bioinformatic analysis of the protein sequences. Indicates the type of experiment as: in vitro, in vivo or yeast-two-hybrid. Link to GenProt Viewer (makes it possible to visualize the protein information in the context of the related gene)	-	[24] http://www.hprd.org/	Yes	No
STRING Search Tool for the Retrieval of Interacting Genes/ Proteins	Known and predicted protein-protein interactions derived from 4 sources: predictions based on genomic context analysis, high-throughput experimental data, co-expression experiments, and mining of databases and the literature	Available	[28] http://string.embl.de/	Yes	Yes
MINT Molecular Interaction	Experimentally verified protein interactions mined from scientific literature by expert curators HomoMINT contains inferred interactions between human proteins. Contains interactions involving non-protein entities such as promoter regions and mRNA transcripts Gives information on kinetics, binding constants and interaction of participating domains Separate annotation of human protein interactions in HomoMINT.	Available through the MINT Viewer Data can be exported to Osprey	[37] http://mint.bio.uniroma2.it/mint/Welcome.do	Yes	No

Language Processing (NLP) is an algorithm used for the automated mining of abstracts and titles from Pubmed articles [38]. The information extraction technology was used to search sentences in Medline abstracts that support previously known interactions annotated in the DIP database. Correspondence between DIP protein pairs and Medline sentences describing their interactions was found in only 30% of the cases [39]. Nevertheless, the information extraction system has the ability to identify new relations between proteins [39, 40]. Once again this example illustrates the importance of using information contained in published papers.

A NLP-based text-mining approach is Chilibot (chip literature robot). This software queries the Pubmed database and retrieves content-rich relationship networks among biological concepts, genes, proteins, or drugs [41]. An additional advantage of this web tool is that of the characterization of each molecular relationship as inhibition or stimulation, and based on directionality (Figure 4).

Another tool for retrieving the information contained in Pubmed is iHop (Information Hyperlinked over Proteins), a literature network developed using genes and proteins as hyperlinks between sentences and abstracts [42]. Besides literature information, iHop also contains interactions collected from external resources (*e.g.* large-scale experimental data), protein homologues and links to external resources (*e.g.* UniProt, NCBI, OMIM) [42]. The visualization system is user-friendly because within the references, which are ranked according to their significance, important expressions such as diseases, biological processes or other proteins, are highlighted in different colors.

Text mining algorithms have been improved using field specific synonym dictionaries, longer word strings for search, full text articles for queries and statistical methods [38]. Nevertheless, although the approach accelerates the discovery of relevant information, the reliability of the results is less than that achieved by a curator who examines each paper [38].

Biological pathways: Building blocks of functions

The integration of biochemical properties of proteins is represented in biological pathways. Therefore, in a post-proteomic study, pathway analysis is an essential step for the systematic understanding of cellular activities.

About 240 biological pathway resources are contained in the meta-database Pathguide [43, 44]. Some metabolic and signaling pathways databases are iPath, Reactome, and Protein Lounge (Table 3). iPath is an online resource with only hundreds of human pathways. But Reactome, in addition to curated human events, has inferred orthologous

31. Jayapandian M, Chapman A, Tarcea VG, Yu C, Elkiss A, Ianni A, *et al*. Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. Nucleic Acids Res 2007;35:D566-71.

32. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34: D535-9.

33. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucl Acids Res 2004;32:D449-51.

34. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, *et al*. The Biomolecular Interaction Network Database and related tools 2005 update. Nucl Acids Res 2005;33:D418-24.

35. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C et. al. IntAct-open source resource for molecular interaction data. Nucleic Acids Res 2007;35: D561-5.

36. Mewes HW, Frishman D, Mayer KFX, Munsterkotter M, Noubibou O, Pagel P et. al. MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucl Acids Res 2006;34:D169-72.

37. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS Lett 2002;513(1): 135-40.

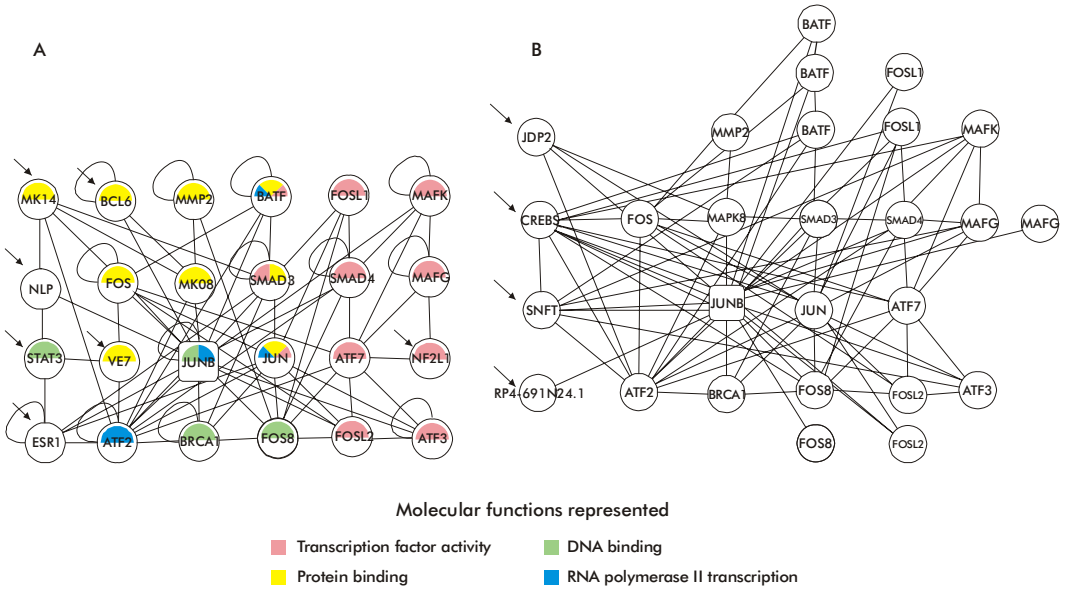

Figure 3. Protein-protein interaction network constructed in Cytoscape with the plugins: A-) APID2NET and B-) MIMI. In the network created with APID2NET (A) the nodes are highlighted according to the molecular function annotated in the Gene Ontology database. This functionality is not available with the MIMI plugin. Although most of the interactions are common in APID and MIMI, some interactions are only annotated in one of these integrative databases (represented with arrows on the nodes)

events in 22 non-human species, including *Mus musculus, Rattus novergicus, Escherichia coli, D. melanogaster* and *S. cereviseae* [45]. Moreover Reactome can be downloaded and has a tool, named: Skypainter, which is useful for determining statistically over-represented events (reactions and/or pathways) in a set of genes or protein identifiers [46]. On the other hand, Protein Lounge is a commercial package that is not available for the academic community. A similar alternative is the free database BioCarta (Table 3), which has useful information to understand the biological processes represented in the pathway maps.

Perhaps one of the most popular pathway databases is provided by the Kyoto Encyclopedia of Genes and Genomes (KEGG). KEGG is an integrated resource with four main databases categorized as: building blocks in the genomic space (Gene database) and in the chemical space (Ligand database), diagrams in the network space (Pathway database) and ontologies for pathway reconstruction (Brite database) [47]. Specifically, the Pathway database is a collection of manually drawn diagrams called the KEGG reference pathway diagrams (maps) [48] (Table 3). With these reference maps organism-specific pathways are automatically generated by coloring the genes of the given organisms [48].

Another repository of pathways involved in both primary and secondary metabolism is MetaCyc (Table 3). This database also stores review-level comments as well as enzyme information, which include substrate specificity, kinetic properties, activators, inhibitors, co-factor requirements and links to sequence and structure databases [49]. Most of the pathways described in MetaCyc occur in microorganisms and plants, although animal pathways are also represented [50].

Cellular behavior and organization are determined by the "cross-talk" among different pathways. In this sense, visualizing all the pathways associated with the experimental data in one framework facilitates the study of biological systems. Usually this type of analysis is not available through pathway databases. Software on biological networks are recommended to overcome this limitation.

Biological networks: interrelating pathways

Different types of biological networks, such as: protein-protein interaction, metabolic or gene regulatory networks can be reconstructed based on previous knowledge. In general, biological networks have

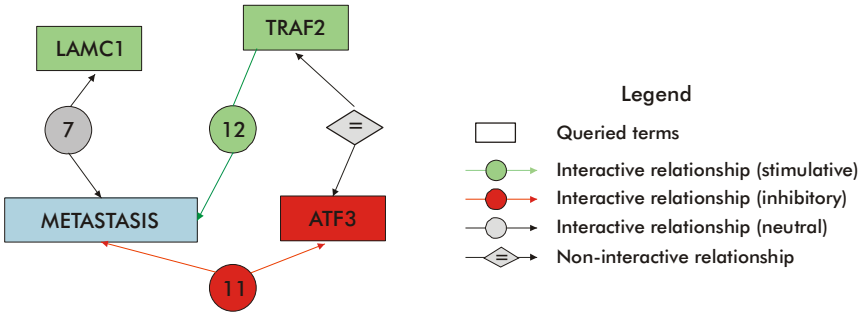

Figure 4. Biological map constructed by Chilibot. Chilibot queried the entire PubMed abstract database to identify relationships between a biological concept («metastasis») and a set of genes reported to be up- (TRAF2-TNF receptor-associated factor 2-, LAMC1-Laminin subunit gamma 1-) and down-regulated (ATF3-Cyclic AMP-dependent transcription factor ATF-3-) in metastatic prostate cancer compared to localized prostate cancer [51]. Papers retrieved indicate that TRAF2 stimulates the process of metastasis, while ATF3 has an inhibitory relationship. The number within the icon of each line indicates the number of abstracts retrieved that document that relationship. The arrows indicate the direction of the interaction. The colors green or red of the rectangular nodes represent up- or down-regulation of the genes, respectively, based on experimental data provided by the user. Nodes with no expression values (e.g., metastasis) are in cyan. The terms and icons are linked to documentation when viewed in a web-browser.

38. Chaussabel D. Biomedical literature mining: challenges and solutions in the 'omics' era. American Journal of Pharmacogenomics 2002;4:383-93.

39. Blaschke C, Valencia A. Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. Comparative and Functional Genomics 2001;2:196-206.

40. Santos C, Eggle D, States D. Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. Bioinformatics 2005;21:1653-8.

Table 3. Databases of biological pathways.

Database name	Description	Reference URL	A_V	C_V
iPath	Collection of 225 human biological signaling and metabolic pathways covering over 2 700 distinct genes/proteins	http://www.invitrogen.com/ipath	Yes	No
REACTOME	Database with human biological processes curated from the literature, includes significant comments and literature citations	[45] http://www.reactome.org/	Yes	No
Protein Lounge	Contain the Pathway Database, a collection of metabolic and signaling pathways available for many organisms. References and information on the pathways and the related proteins are included in the database	http://www.proteinlounge.com/	No	Yes
BioCarta	Maps and summarizes biological processes. The online maps are dynamic graphical models that provide information for over 120000 genes from multiple species	http://www.biocarta.com/genes/index.asp	Yes	No
KEGG Pathway	Contains pathways maps of metabolism, genetic information processing, environmental information processing, cellular processes, human diseases and drug development	[47] http://www.genome.jp/kegg/	Yes	No
MetsCyc	Contains experimentally verified pathways and enzyme information curated from the scientific literature	[50] http://metacyc.org	Yes	No

DB: Databases, A_V: Academic Version, C_V: Commercial Version

emergent properties generated by the interaction of their components. They also have a scale-free and modular organization. In scale-free networks the degree of distribution (number of links per node) follows a power-law, which means that only a small number of nodes, called hubs, are highly connected [52]. Hubs usually play essential roles in biological systems [53]. On the other hand, groups of proteins with similar functions tend to form clusters or modules in the network architecture [54]. These structural features contribute to the robustness of the biological system and may explain the fact that many drug candidates are ineffective or show unexpected severe side effects. At the same time it highlights why a post-proteomic study data must be also analyzed from a network perspective.

Ingenuity Pathway Analysis, MetaCore and PathwayStudio are commercial software designed to visualize high-throughput data in the context of biological networks (Table 4). In addition, Ingenuity Pathway Analysis and MetaCore identify the most relevant functions and pathways represented in experimental gene or proteins datasets. Both applications make use of manually curated databases, Ingenuity Pathways Knowledge Base and MetaBase respectively (Table 4). The former is a mammalian network database while the latter has only human

41. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics 2004;5:147.

42. Hoffmann R, Valencia A. A gene network for navigating the literature. Nat Genet 2004;36:664.

43. Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic Acids Res 2006;34:D504-6.

Table 4. Software for network analysis.

Software	Description	Reference URL	A_V	C_V
Ingenuity Pathway Analysis	Based on Ingenuity Pathways Knowledge Base, database of biological networks created from millions of relationships, between proteins, genes, complexes, cells, tissues, drugs, and diseases, extracted manually from the literature Web access. Developed by Ingenuity Systems Inc	http://www.ingenuity.com/	No	Yes
MetaCore	Integrated software suite based on the content of MetaBase, a manually curated database of human protein-protein and protein-DNA interactions, transcriptional factors, signaling, metabolism and bioactive molecules Includes intuitive tools to visualize and exchange data, multiple networking algorithms and in silico filters Web access and in-house installation Developed by GeneGo Inc	http://www.genego.com/	No	Yes
PathwayStudio	Comes with ResNet 5 Mammalian database and ResNet Plant database. These databases are a collection of eukaryotic molecular interactions generated by MedScan Text- to-Knowledge Suite using the entire PubMed database and 43 full text journals. Also works with public databases of signaling and biochemical pathways, including KEGG, BIND and HPRD Desktop product Developed by Ariadne Genomics Inc	http://www.ariadnegenomics.com/	No	Yes
GenMAPP Gene Map Annotator and Pathway Profiler	Archived Maps were drawn based on textbooks, articles and public pathway databases or generated from the public database maintained by the Gene Ontology Project. Maintains data relevant to various species such as: S cerevisiae, C. elegans, R. novergicus, and *H. sapiens*	[55] http://www.genmapp.org/	Yes	No
Cytoscape	Software for integrating biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework	[56] http://www.cytoscape.org/	Yes	No

A_V: Academic Version, C_V: Commercial Version)

information. Also PathwayStudio (formerly known as PathwayAssist) is based on a mammalian database, named: ResNet 5 Mammalian, but in this case it is generated by the text mining of the Pubmed database and 43 full text journals (Table 4) [57]. Another version of ResNet, was simultaneously developed in the same way, and specifically for plants. With PathwayStudio it is possible to draw pathway diagrams and automatically update pathways with newly published facts. In general these tools are useful to select drug targets, validate and identify molecular biomarkers for disease conditions, and propose alternative indications of approved drugs.

Free software is also available in this field. For example Pathway Voyager, a flexible approach that uses the KEGG database to pathway mapping, with just a few prerequisites, it does not require any specific hardware (*i.e.*, a background server) or software (*i.e.*, relational database backbones) [58]. GenMapp also was designed to visualize gene expression data on maps representing biological pathways and gene groupings (Table 4). However GenMapp has more options than Pathway Voyager since users of this software can modify or design new pathways and apply complex criteria for viewing gene expression data on pathways [55].

An additional option is Cytoscape (Table 4), the core of this software provides basic functionality to layout and query the networks, to visually integrate the networks with expression profiles and phenotypes, and to link the networks with databases of functional annotations [56]. Furthermore, a variety of external methods, implemented as plugins, can be used to construct and analyze the networks [59]. For example, besides the previously mentioned plugins APID2NET and MIMI (Figure 3), another resource is a web service client that downloads interaction data from databases such as Intact, Pathway Commons and NCBI Entrez Gene. To analyze the topological parameters of a network, several plugins such as CentiScape and Network Analyzer are available. Also, the network modular organization can be determined with the NetMach and MCODE plugins. The fact that it is an open source has determined the development and improvement of Cytoscape, which is now one of the most useful software for data analysis.

Limitations and future developments

Until 2006, there were 1357 unique therapeutic drugs approved; of which 1065, had a known mode of action, whose action was through only 324 molecular targets [60]. Historically, due to the tendency to re-use the same targets and recognized mechanisms the rate of target innovation (the rate at which drugs against new targets are launched) has been constant [60]. But the improvement in technologies and sequenced genomes are potential advantages to change this trend. The human genome contains 30 000-35 000 genes with the potential to synthesize more than 100 000 proteins [61]. Less than 50% of thes genes can be assigned a putative biological function on the basis of the sequence data [61]. In this scenario the challenge is to determine which proteins might be a viable research target, with estimates about the number of possible targets ranging from 600-1500 [62] to 5000-10 000 [63]. In fact today, the number of targets available is no longer rate limiting, actually the real problem is how to select the targets that are more likely to succeed.

The gene→protein→target→hit paradigm of drug discovery is now recognized as oversimplified, due to the complexity of biological systems. Creating an inventory of genes, proteins and metabolites is necessary but not sufficient to understand the integrated roles of genomes, transcriptomes, proteomes and metabolomes [64]. In any "post-omic" analysis two essential concepts must be applied to understand biological functions at a system level. First, integrate different levels of information and second, view cells in terms of their underlying network structure.

The information on a biological entity is distributed in different databases. Hence, the information retrieval process from diverse sources is time consuming. Moreover, current databases are good for the analysis of a particular protein or small interaction networks, but they are not as useful for the integration of complex information on cellular regulation, pathways, networks, cellular roles and clinical data, and they lack coordination and the ability to exchange information between multiple data sources [65]. At the same time, there is still a need for software that can integrate a wide array of biological attributes, including molecular interactions drawn from the literature and experimental datasets. Therefore, one necessary step for post-proteomic data analysis will be the development of bioinformatic tools to retrieve, manage and integrate significant biological information from different databases.

To analyze data from a network perspective, software like Ingenuity Pathways Analysis, MetaCore and PathwayStudio, which work with owner curated databases, have been created; although their high prices make them practically unaffordable to the academic community. A useful alternative is Cytoscape, which, besides being a tool for the visualization of high-throughput expression data over networks, it has functions that analyze properties and the organization of those networks. However, the networks constructed with Cytoscape are sometimes liable to show errors. In this sense efforts must be focused on improving the quality of the available curated databases and also to develop integrative knowledge bases that are especially designed to construct biological networks. Moreover, in databases it should be mandatory to annotate only high quality experimental results, a concern especially associated with high-throughput outcomes.

The metabolic network is a reliable source of information, a point for the integration of data from genomic, proteomic and metabolomic studies and, moreover, a direct graph in which the down-stream effects of a perturbation can be predicted. Although the metabolic network has important features for drug discovery, its use in this sense, at least in humans, is very limited [66]. Pathway Voyager and GenMapp are two platforms used to map data in metabolic pathways principally derived from the KEGG database. But a human metabolic network was not available until recently when two high quality maps were developed: EHMN (Edinburgh human metabolic network) [67] and

44. http://www.pathguide.org.

45. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, *et al*. Reactome: a knowledgebase of biological pathways. Nucl Acids Res 2005;33: D428-32.

46. http://www.reactome.org/cgi-bin/skypainter2?DB=gk_current.

47. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, *et al*. From genomics to chemical genomics: new developments in KEGG. Nucl Acids Res 2006;34:D354-7.

48. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucl Acids Res 2004;32:D277-80.

49. Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, *et al*. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucl Acids Res 2006;34:D511-6.

50. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, *et al*. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucl Acids Res 2004;32:D438-42.

51. Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, *et al*. The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature 2002;419:624-9.

52. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5: 101-13.

53. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, *et al*. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 2004;430:88-93.

54. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999;402:C47-52.

55. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 2002;31:19-20.

56. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al*. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 2003;13:2498-504.

57. Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics 2004;20:604-11.

58. Altermann E, Klaenhammer TR. PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. BMC Genomics 2005;6:60.

59. http://www.cytoscape.org/plugins2.php.

60. Overington JP, Al Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov 2006;5:993-6.

61. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al*. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.

Recon1 [68]. The combination of the two networks can generate a more complete and reliable map. However, because of the different compound names used in the two networks, they are not easily merged [66]. Additional efforts will also be needed for metabolic network reconstruction and analysis.

The progress in technologies and the sequenced genomes offer opportunities to study genes and proteins on a larger scale than ever before. In particular, proteomics may yield crucial information on the regulation of biological functions and the mechanism of diseases. In this sense it is a highly promising area for drug discovery. But further advances in bioinformatics are critical, not only for the interpretation of large data sets but also to integrate data from different biological levels.

Acknowledgements

We thank Alexis Mussachio Lasa and Luis Javier González for their helpful comments and sugges-tions.

62. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov 2002;1: 727-30.

63. Drews J. Drug discovery: a historical perspective. Science 2000;287:1960-4.

64. Kitano H. Systems biology: a brief overview. Science 2002;295:1662-4.

65. Tucker CL, Gera JF, Uetz P. Towards an understanding of complex protein networks. Trends in Cell Biology 2001;11:102-6.

66. Ma H, Goryanin I. Human metabolic network reconstruction and its impact on drug discovery and development. Drug Discov Today 2008;13:402-8.

67. Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, *et al*. The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol 2007;3:135.

68. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, *et al*. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 2007;104: 1777-82.

Received in october, 2007. Accepted for publication in september, 2008.