Review **Open Access**

Angelo Facchiano*

Bioinformatic resources for the investigation of proteins and proteomes

DOI 10.1515/ped-2017-0001
Received October 5, 2015; accepted January 22, 2016

Abstract: Experimental techniques in omics sciences need strong support of bioinformatics tools for the data management, analysis and interpretation. Scientific community develops continuously new databases and tools. They make it possible the comparison of new experimental data with the existing ones, to gain new knowledge. Bioinformatics assists proteomics scientists for protein identification from experimental data, management of the huge data produced, investigation of molecular mechanisms of protein functions, their roles in biochemical pathways, and functional interpretation of biological processes. This article introduces the main bioinformatics resources for investigation in the protein world, with references to analyses performed by means of free tools available on the net.

Keywords: Databases, bioinformatics tools, experimental interpretation, computational tools, Gene Ontology

1 Introduction – Bioinformatics from origins to now

Bioinformatics is a field of research born in the 1960s, as a natural consequence of the simultaneous growth of computational power and of experimental data. Protein three-dimensional structures, together with amino acid and nucleotide sequences, suggested scientists to collect such information within ordered archives, and to generate software able to manage the data and elaborate new information. Scientists of the protein world are indebted to the pioneering work of Margaret Dayhoff in this field [1]. Computers, fifty years ago were only present in few laboratories. They automatized the procedures for database management and the analysis and comparison of sequences for evolutionary studies. For some decades, the main tasks for bioinformatics have been considered management of archives, comparison of sequences to identify functional and evolution relationships, and molecular structure visualization. At the end of the second millennium, the role of bioinformatics appeared crucial for the emerging omics world. The use of bioinformatics in assembling genome sequence information has been decisive to reach the conclusion of the Human (and other) Genome Projects. The large amount of data generated by each omics approach demonstrated to the research community that bioinformatics tools and expertise were definitely crucial to the success of modern research in life sciences. Therefore, new interest for bioinformatics has started during the last twenty years.

Omics sciences require the management of large amount of data, at a first step, but also to analyze them massively, and compare them also when they come from different technical methods. For a complete view of the results of any omics study, a large use of bioinformatics is an obligate step. This has given a strong impulse to bioinformatics studies, and at the same time, the development of novel tools has been required by the continuous issue of novel techniques to sequence, identify and investigate molecules by high-throughput approaches. Bioinformatics also evidently needs standardization, at all levels of the research workflow, in order to produce experimental data that are comparable, also from different technical approaches, in view of integrating more results and obtaining high-quality analyses. Standardization of the data presentation and deposition in archives is also a well-known task in omics sciences [2-3] and many efforts to standardization and data sharing are ongoing [4-5]. Therefore, bioinformatics has taken on the role of inspector of quality of experimental data, in addition to their analysis and interpretation. Fig. 1 schematizes the role of bioinformatics in receiving and giving back information to omics sciences applied to the study of biological systems, creating relationships among them, and driving the interpretation of the biological phenomena under investigation.

***Corresponding author Angelo Facchiano,** National Research Council, Institute of Food Science, Laboratory of Bioinformatics, via Roma 64, 83100 Avellino, Italy. angelo.facchiano@isa.cnr.it

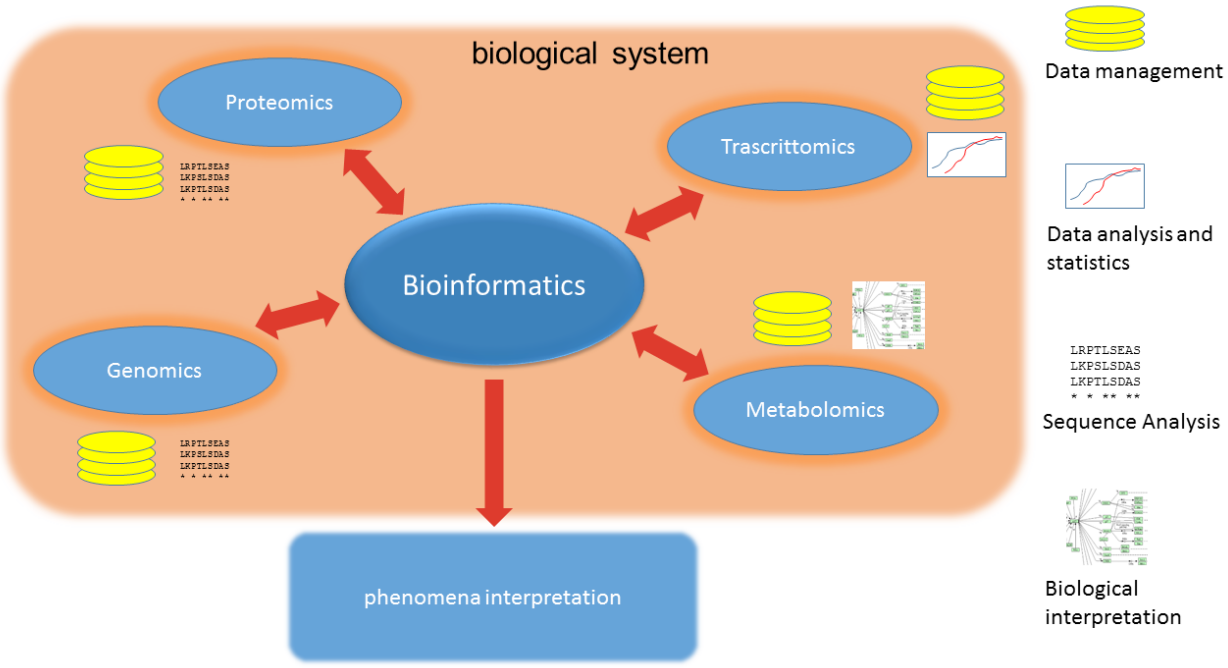

Figure 1: The scheme indicates the central role of bioinformatics in the modern biological investigation based on omics sciences. Legend for small images is on the right side of the figure. They indicate the main role assumed by bioinformatics for that omics technique, although a wide application can be assumed in all cases.

Peptidome investigation involves bioinformatics firstly into the identification of the amino acid sequence that produces the identified peptides. Moreover, further bioinformatics tools can be used for investigation at functional level. To these aims, many bioinformatics resources are freely available on the net by web interfaces. Catalogs of bioinformatics tools, or review articles, may be useful to find the most interesting, although the simplest way is to refer to scientific literature, and a real updated list of tools can be obtained only by using web search engines. In scientific literature, specialized journals and special issues are regularly published to describe databases, tools and web service applications in bioinformatics. This is the reason for not reporting in this article tables listing them, being a useless and incomplete description of the current resources. The reader can search for the most recent special issues, or surf in the net. Therefore, I find it more useful to discuss the categories of tools with indications of the most common and potential application from protein identification to functional analysis and interpretation of biological phenomena.

The role of bioinformatics in the proteomics field is clear by considering the work on human proteome [6-7] as well as many other large-scale studies (too many to be listed here). Bioinformatics resources consist mainly of two types of software: databases and tools for data analysis. The next paragraphs introduce some of the most popular in both types. A general comment concerns the usage of these tools. Usually, databases are accessed by web interfaces, so that users do not need to manage data bases and searching tools on own hardware. This was a very common task some decades ago, when the internet was still not used to distribute databases. The net makes it possible to use resources on the other side of the world. However, the dark side of the moon hides a crucial aspect in research, i.e. the reproducibility of results. The results of a database search could become obsolete within a week, due to continuous addition of data in archives, which means that differences in results can appear when the search is repeated after some time. Curators of the most common databases publish periodically new releases: monthly, weekly, or daily. The previous releases are sometime still available for downloading, but not searchable by the web interface, so if an article reports results of a BLAST search, during the time needed for article publication, it is likely that the used database has been updated, and the results could change. Similarly, also the computational tools may change over the time. This means that, for reproducibility of results, it should be a good practice to publish the results of database searching with the date of the analysis, and the releases of the database and tool used.

2 Web portals

The term "portal" refers to web sites that introduce to a specific world. In bioinformatics, some portals are considered reference points for access to resources, databases, and tools. The most relevant in the field of proteomics is the Expasy portal (www.expasy.org), active since 1993. In addition, the most relevant sites for all aspects of bioinformatics are the web site of the National Center for Biotechnological Information, NCBI (www.ncbi.nlm.nih.gov) and of the European Bioinformatics Institute, EBI (www.ebi.ac.uk), where many other resources for proteomics can be found.

3 Sequence databases

The reference resource for amino acid sequence is UniProt, a database managed by a consortium created in 2002 [8] by the teams that in the previous decades maintained SwissProt, TREMBL and PIR databases. These three resources, although different in their origins and forms, reported redundant information, and in addition became difficult to maintain for their growth and costs. Therefore, the creation of a consortium among the developers allowed the migration towards a unique resource, named UniProt, aimed to reduce redundancy of databases, and improve the quality of information collected.

The UniProt database consists of different data sets, the most used being UniProt/SwissProt and UniProt/TREMBL sections that substitute the old three databases. UniProt/SwissProt collects amino acid sequences with highly curated annotations, with monthly new releases. The 2015_09 release (the number indicates year and month of release) collects more than 549.000 sequences. UniProt/TREMBL collects sequences obtained by automatic translation of nucleic acid sequences included in EMBL database (a database of nucleic acid sequences), and the 2015_09 release included more than 50 Million sequences. Although the existence of proteins and peptides corresponding to the translation of any nucleic acid sequence is not proven at the experimental level, TREMBL resource represents a useful tool to be used for putative identification of novel sequences and correlation to previous experimental data.

In addition to EMBL, the other main collection of nucleic acid sequences is GenBank [9]: as for the TREMBL collection, it is possible to find other amino acid sequences from the translation of GenBank sequences. The whole collection, from direct or indirect observation, consists of about 70 Million amino acid sequences.

Other sequence databases should be taken into account for their specialization. As an example, the Phytozome database [10] represents an integrated collection of genomic sequences from vegetable species, and in some cases, this database represents the unique source for sequences still waiting for enclosure in Genbank or EMBL databases. Therefore, protein sequences of such species, as coded into the genome, can be obtained from Phytozome database. Similarly, a huge source of sequences is the Genomes Online Database (GOLD) [11] and other genome sequence collections on the net.

Specialized databases can also present small collections, focused on peculiar aspects. As an example, I would like to remember the work that I coauthored to develop of a manually curated collection of active sequences from proteins, named ASC (Active Sequence Collection) [12], focused on short protein subsequences with a recognized biological activity, useful for detecting potentially active regions within the protein of interest. This is an alternative approach not available by online common tools, searching for similarity within whole databases with a lot of redundancy and without direct information for specific biological activities.

4 Bioactive peptides databases

As pointed out in the introduction paragraph, it is not possible to list all resources on the net, and probably it would be not really useful, because in short time some database will be inoperable, due to migration to new address or because discontinued by the authors. However, to give some information useful as a starting point, here I report some example of specialized database in different field.

Some databases of peptides with antimicrobial activity are APD [13], AMPer [14], BaAMPs [15]. Other databases are specialized on allergens, as Immune Epitope Database [16]. Lists of databases of allergens or bioactive peptides are available in literature [17] or by surfing the net. Similarly, for bioactive peptides I just quote the BIOPEP resource [18] that includes subsections for allergens. The web site hosting BIOPEP lists also other bioactive peptides databases (http://www.uwm.edu.pl/biochemia/index.php/pl/biopep/32-bioactive-peptide-databases). The link lists 48 bioactive peptide databases, with the related URL. This list will surely not be exhaustive, as any catalog of bioinformatic resources.

5 Browsing the sequence databases

The usefulness of a database depends on the capability to extract information from it. Sequence databases are collections with two kinds of data: the mandatory information is the amino acid sequence, the other, not less relevant but optional, is any annotation concerning the given protein. Both kinds of information can be used to search within the database and extract in a direct way the protein of interest. However, the searching tools are different for the two kinds of information. While annotations can be searched with a text keyword search, just like a common search engine on the web, amino acid sequences are searched with specific tools, which compare a reference sequence (query) with each sequence within the database. The result is a list of proteins that are similar to the query. The most popular tool to make this search is BLAST, acronym for Best Local Alignment of Sequence Tags [19]. It can be used on different web sites, and this means that results obtained may differ, depending on the standard settings applied and the database configured for searching. Some example will explain these differences and the meaning of the results. BLAST evaluates the similarity of the query with each sequence of the database, by using matrices of similarity among the amino acids, applying penalties for gaps, and the results depend on the choices of these and few other parameters. It is very simple to use BLAST with standard settings, but it is recommended to know what they mean and how they can affect the results. Tables 1, 2 and 3, show the results obtained with BLAST searches, using the human myoglobin sequence as query (UniProt ID code: MYG_HUMAN). In the example 1 (Table 1), BLAST has been used at the NCBI web site, and the non-redundant database (nr) has been selected, which means a collection of about 70 millions of sequences, from GenBank CDS (coding sequences), UniProt/SwissProt, Protein Data Bank (PDB) and others, verified to avoid redundancy of sequences. The differences between examples no. 1 and 2 (Table 2) is the settings of the similarity matrix. The two searches have been performed by modifying the similarity matrix, BLOSUM45 and BLOSUM90, respectively. In both

Table 1: BLAST results, example A. See the text for explanations.

Example A - web site: NCBI BLAST; db: nr; matrix: BLOSUM45

Description	Max score	Total score	Query cover	E value	Identity	Accession
myoglobin [Homo sapiens]	349	349	1	2E-112	1	NP_005359.1
myoglobin [synthetic construct]	349	349	1	2E-112	1	AAX36993.1
myoglobin, isoform CRA_a [Homo sapiens]	350	350	1	4E-112	1	EAW60065.1
myoglobin transcript variant 1 [Homo sapiens]	347	347	1	6E-112	0.99	AAX84516.1
myoglobin [Homo sapiens]	347	347	1	9E-112	0.99	AAA59595.1
PREDICTED: myoglobin isoform X2 [Nomascus leucogenys]	347	347	1	1E-111	0.99	XP_003264735.1
PREDICTED: myoglobin isoform X2 [Pan troglodytes]	346	346	1	2E-111	0.99	XP_001156696.1
unnamed protein product [Homo sapiens]	346	346	1	3E-111	0.99	BAG36764.1
PREDICTED: myoglobin isoform 1 [Gorilla gorilla gorilla]	346	346	1	3E-111	0.99	XP_004063429.1
PREDICTED: myoglobin isoform X1 [Pan troglodytes]	347	347	1	4E-111	0.99	XP_009436537.1
PREDICTED: myoglobin isoform X1 [Pan paniscus]	347	347	1	4E-111	0.99	XP_008973239.1
Myoglobin	345	345	0.99	9E-111	0.99	711658B
myoglobin [Pongo abelii]	344	344	1	2E-110	0.99	NP_001125556.1
PREDICTED: myoglobin isoform X1 [Nomascus leucogenys]	349	349	1	4E-110	0.99	XP_003264736.1
Myoglobin	342	342	0.99	6E-110	0.98	711658A
Chain A, Crystal Structure Of Human Myoglobin Mutant K45r	342	342	0.99	1E-109	0.99	3RGK_A
Myoglobin	341	341	0.99	3E-109	0.99	761377A
PREDICTED: myoglobin isoform X1 [Colobus angolensis palliatus]	335	335	1	4E-107	0.96	XP_011816533.1
PREDICTED: myoglobin [Papio anubis]	335	335	1	7E-107	0.96	XP_003905518.1
PREDICTED: myoglobin isoform 3 [Macaca mulatta]	333	333	1	3E-106	0.95	XP_001082215.1

Table 2: BLAST results, example B. See the text for explanations.

Example B - web site: NCBI BLAST; db: nr; matrix: BLOSUM90

Description	Max score	Total score	Query cover	E value	Ident	Accession
myoglobin [Homo sapiens]	328	328	1	2E-113	1	NP_005359.1
myoglobin, isoform CRA_a [Homo sapiens]	328	328	1	2E-113	1	EAW60065.1
myoglobin [synthetic construct]	328	328	1	2E-113	1	AAX36993.1
myoglobin transcript variant 1 [Homo sapiens]	327	327	1	6E-113	0.99	AAX84516.1
PREDICTED: myoglobin isoform X2 [Nomascus leucogenys]	326	326	1	9E-113	0.99	XP_003264735.1
myoglobin [Homo sapiens]	326	326	1	9E-113	0.99	AAA59595.1
PREDICTED: myoglobin isoform X1 [Nomascus leucogenys]	328	328	1	1E-112	0.99	XP_003264736.1
PREDICTED: myoglobin isoform X2 [Pan troglodytes]	325	325	1	2E-112	0.99	XP_001156696.1
PREDICTED: myoglobin isoform X1 [Pan troglodytes]	326	326	1	3E-112	0.99	XP_009436537.1
PREDICTED: myoglobin isoform 1 [Gorilla gorilla gorilla]	325	325	1	3E-112	0.99	XP_004063429.1
PREDICTED: myoglobin isoform X1 [Pan paniscus]	325	325	1	4E-112	0.99	XP_008973239.1
unnamed protein product [Homo sapiens]	325	325	1	4E-112	0.99	BAG36764.1
Myoglobin	325	325	0.99	5E-112	0.99	711658B
Myoglobin	323	323	0.99	2E-111	0.98	711658A
myoglobin [Pongo abelii]	323	323	1	2E-111	0.99	NP_001125556.1
Chain A, Crystal Structure Of Human Myoglobin Mutant K45r	322	322	0.99	7E-111	0.99	3RGK_A
Myoglobin	320	320	0.99	2E-110	0.99	761377A
PREDICTED: myoglobin isoform X1 [Colobus angolensis palliatus]	317	317	1	6E-109	0.96	XP_011816533.1
PREDICTED: myoglobin [Papio anubis]	315	315	1	2E-108	0.96	XP_003905518.1
PREDICTED: myoglobin isoform 3 [Macaca mulatta]	314	314	1	7E-108	0.95	XP_001082215

cases, the human myoglobin has been recognized, being a well-known protein whose sequence is present in the database searched. In these cases, the differences in the parameter settings can be quite ineffective at the results level. In fact, subtle differences in the two examples in terms of scores and E-values, to be ascribed to the different settings of similarity matrix, do not affect the main scope of the search, that is to identify the protein corresponding to the query sequence. In addition, some difference appears in the order of results, sorted by the computed values, although the best results are the same and clearly identify the query as the human myoglobin.

In the example no. 3 (Table 3) the same query sequence has been searched by means of BLAST at the UniProt web site, where the standard settings for database is the UniProtKB. Also in this case, the best results refer to entries of the human myoglobin, so also this search identifies correctly the query. Results are still different from the previous examples in terms of scores and e-values, and accession numbers and protein names appear different (as expected, due to the different database used). One could wonder what is the usefulness of these values, what is the correct result of the BLAST search, or in other terms, what of the three searches is the correct one. The answer is that each of the examples gives correct result, as protein identification, and as values, because they result from a computation based on the settings. These values help the user to evaluate the quality of the query identification. The exact values of scores and e-values are indicative and depend on the database used and the parameters of the search.

Two type of similarity matrices can be selected in these searches, i.e. PAM and BLOSUM, differing for their origin. In both cases, they evaluate the similarity of amino acids on the basis how they substitute each other during evolution. PAM matrices, from the work of Margaret Dayhoff, have been generated by the alignment of homologous protein sequences, while BLOSUM from the alignment of conserved segments, collected in BLOCKS database [20]. The name of each PAM and BLOSUM matrix includes a numerical suffix, used with opposite significate. In comparisons of distantly related sequences,

Table 3: BLAST results, example C. See the text for explanations.

Example C - web site: UniProt; db: UniProtKB; matrix: BLOSUM62

Description	E-value	Score	Identity	Accession
MYG_HUMAN - Myoglobin Homo sapiens (Human)	5.70E-108	809	100.00%	P02144
A0A024R1G3_HUMAN - Myoglobin Homo sapiens (Human)	8.60E-108	809	100.00%	A0A024R1G3
MYG_PANTR - Myoglobin - Pan troglodytes ...	3.30E-107	804	99.40%	P02145
G1RW45_NOMLE - Myoglobin - Nomascus leucoge...	4.70E-107	803	99.40%	G1RW45
MYG_SYMSY - Myoglobin - Symphalangus syn...	4.70E-107	803	99.40%	P62735
MYG_HYLAG - Myoglobin - Hylobates agilis...	4.70E-107	803	99.40%	P62734
G3R764_GORGO - Myoglobin - Gorilla gorilla ...	9.40E-107	801	99.40%	G3R764
MYG_GORBE - Myoglobin - Gorilla gorilla ...	9.40E-107	801	99.40%	P02147
B2RA67_HUMAN - Myoglobin Homo sapiens (Human)	2.70E-106	798	98.70%	B2RA67
MYG_PONPY - Myoglobin - Pongo pygmaeus ...	1.50E-105	793	98.70%	P02148
A0A0D9R744_CHLSB - Myoglobin - Chlorocebus saba...	1.70E-102	773	96.10%	A0A0D9R744
MYG_SEMEN - Myoglobin - Semnopithecus en...	1.70E-102	773	96.10%	P68085
MYG_PAPAN - Myoglobin - Papio anubis (Ol...	1.70E-102	773	96.10%	P68084
MYG_ERYPA - Myoglobin - Erythrocebus pat...	1.70E-102	773	96.10%	P68086
G7PFA9_MACFA - Myoglobin - Macaca fascicula...	6.80E-102	769	95.50%	G7PFA9
F7ARV1_MACMU - Myoglobin - Macaca mulatta ...	6.80E-102	769	95.50%	F7ARV1
MYG_MACFA - Myoglobin - Macaca fascicula...	6.80E-102	769	95.50%	P02150
L5KID3_PTEAL - Myoglobin - Pteropus alecto ...	2.80E-101	765	94.80%	L5KID3
MYG_ROUAE - Myoglobin - Rousettus aegypt...	7.90E-101	762	94.20%	P02163
MYG_PIG - Myoglobin Sus scrofa (Pig)	3.20E-100	758	93.50%	P02189

the most indicated matrices for evaluation of similarity are PAM with high number (as an example, PAM100) or BLOSUM with low number (BLOSUM45). The opposite is suggested for comparisons of closely related sequences.

In the shown examples, the query sequence is present in the database, so it is found and clearly identified. What happens in case of novel sequences, not present in the searched database? The results will report the most similar sequences in the database, but not equal to the query. In case of high similarity (> 70%) and large query coverage, the query can be identified as a homolog of the found sequence(s), scores and e-values will be very high and very low, respectively, and the result is quite simple to be interpreted. In case of very low similarity, some care is required to improve the search and obtain the best possible result. The degree of sequence similarity should be evaluated in terms of percentage of identity and query coverage. If these values are low, they indicate evolutive distance among the query sequence and the sequences found in the database, so it is suggested to repeat the search by using the appropriate settings, and use for own evaluations scores and e-values obtained with the most suitable matrices.

6 Protein 3D Structure Databases

Protein Data Bank (PDB) collects three-dimensional structures of biomolecules, mainly intended as proteins, although nucleic acids are also present both as single molecules and associated to proteins. PDB includes experimental structures, obtained mainly by X-ray crystallography and, at a minor extent, by NMR, while other techniques are also represented by a few structures at low resolution (i.e. > 4 Å). Small peptides are mainly characterized by NMR, technique suitable for molecules that do not crystallize, while X-ray is used if the peptide crystallizes as molecule alone or in association to a larger molecule. A simple analysis on about 300 structures of peptides with sequence length < 16 indicates that NMR has been used for about 96% of them.

Other databases exists for investigation at the three-dimensional structure level. While the primary source of structural data is PDB, there are databases derived from PDB that offer more advanced information. CATH and SCOPe, as an example, are resources useful to investigate the structural organization at level of class and architecture, as well as of protein family. PDBsum,

on the other side, offers summaries of the experimental and structural information related to each PDB file. To reduce the redundancy in the PDB content and select representative structures only, PDB-SELECT is another derived database that, after analysis of the amino acid sequences of PDB files and information about resolution and quality, offers lists of PDB codes that select the most representative protein structures, filtered with different criteria.

7 Proteomics databases

PRIDE (www.ebi.ac.uk/pride/) is a database maintained at EBI for collecting mass spectrometry proteomics data [21]. Since 2004, it is a repository of data concerning protein and peptide identifications ad quantitative values, including post-translational modifications. Data are submitted to PRIDE by means of the ProteomeXchange consortium (http://proteomecentral.proteomexchange.org) that provides a single point for submitting mass spectrometry based proteomics data to public-domain repositories. After submission, datasets are handled by expert curators, which means that it has an "added value" in respect to simple repositories of data without a checkpoint for quality and presence of correct annotations.

Another resource named PEPTIDOME was developed at the NCBI, similarly aimed to collect proteomics experimental data. However, after some years of activity, the database has been discontinued, due to funding constraints, and its resources converged on PRIDE [22].

Another resource is represented by the World-2DPAGE Constellation project [23] that includes many 2D-PAGE resources as experimental 2D-PAGE results and software tools for assisting researchers working in this area.

8 Tools for sequence identification

Researchers involved in proteomics analysis and peptides identification know the most popular resources for mass spectrometry data analysis, as Mascot, ProteinProspector, Sequest, and so on. These tools are oriented to mass spectrometry specialists that become confident with their use in short time. A large number of other bioinformatics tools are available on the net, and lists of links can be found in web portals or specialized sites, as at the Expasy web site (www.expasy.org). These tools may be used to calculate useful properties as expected molecular weight, isoelectric point, peptide fragmentation, or to analyze mass spectrometry data, and for many other applications.

9 Tools for functional analysis

Many bioinformatics tools for different kind of studies can investigate amino acid sequences and three-dimensional structures of proteins to predict functions, by amino acid characteristics or by molecular simulations. These tools require high expertise in structural biology and chemistry. From a strictly bioinformatics point of view, other functional analyses have been developed and offer very interesting applications. Gene Ontology (GO) tools represent one of the most interesting group of bioinformatics resources. Although the term "gene" seems to pone distance from the protein level of investigation, the Gene Ontology is a collection of terms related to three categories, i.e. molecular function, cellular localization, and biological process, that are related each other by logical connections, when applicable. As an example, the molecular function "calcium binding" term is connected with the more general "metal ion binding" term. Furthermore, the terms "cellular component" and "nucleus" are related, as "cellular component" and "membrane", while "nucleus" and "membrane" are not directly related. Examples of the relationships among terms are in Figs. 2 and 3. They have been obtained by using QuickGO [24], one of the tools on the net to browse the content of the controlled vocabulary. The "outer mitochondrial membrane protein" Gene Ontology term is related to other terms by means of the two most frequent relations, i.e. the "is a" and "part of" (Fig. 2). In the example of the term "Negative regulation of autophagy" (Fig. 3) other relations are also evidenced, concerning positive and negative regulation.

The whole number of terms, and their relationships, constitutes the Gene Ontology, i.e., an ordered representation of the knowledge about the genes, and consequently about the product of their expression. Starting from a list of proteins, or genes, Gene Ontology tools help to find their relationships, in terms of molecular functions, or cellular localizations, of biological processes. As an example, a proteomics experiment generates a list of proteins that are differentially expressed between a sample and a reference state. Gene Ontology tools extract the terms related to the list of proteins, and analyze them for enrichment in specific functions, or relationships to metabolic pathways, or to pathologies, with statistical measurements of the significance. This is possible because Gene Ontology tools are advanced bioinformatics software that well integrate many different level of information, and link directly to other advanced resources on metabolic pathways, pathologies, available as external resources, i.e. KEGG database [25], OMIM [26] and so on.

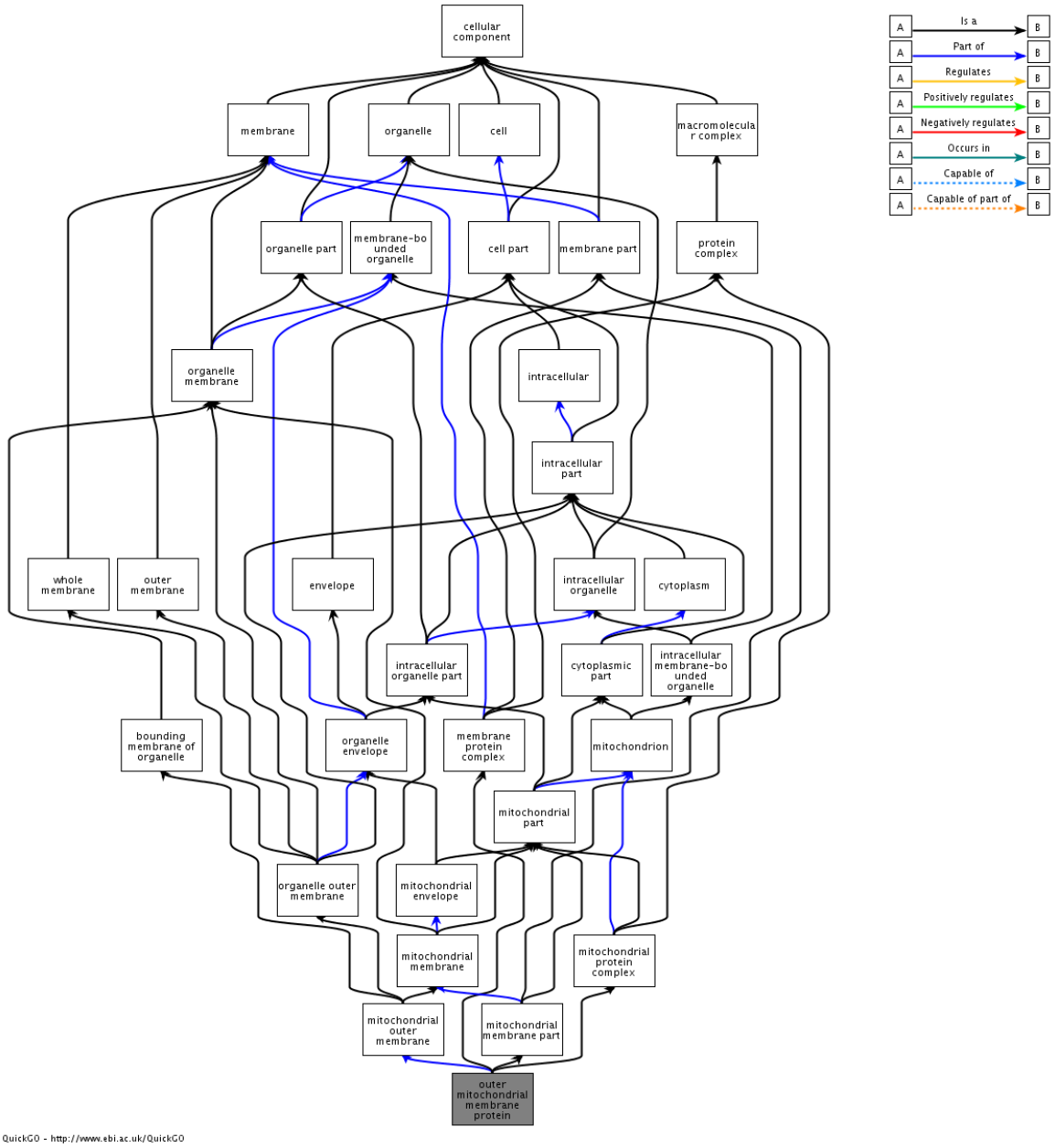

Figure 2: Relations of "outer mitochondrial membrane protein" Gene Ontology term with others. The image is from the QuickGo server [Binns et al., 2009]. Legend explains the possible relations among terms.

10 Bioinformatics' weak points

Anyone can list problems encountered during the own experience when approaching bioinformatics. In my point of view, the main weak point of bioinformatics is the needs of standardization and consolidation of services. Bioinformatics has been useful to evidence this criticism in omics sciences, but is itself affected by the lack of standardization. A typical problem occurring is related to the continuous development of novel functions and novel releases of the tools. It may be needed to reuse a web tool, sometime after its use for the analysis of a data set, but the original web tool is no more available and substituted by a new release.

It could be also difficult to refer to exact web addresses, by considering that during the years web sites may change name, or domain. This occurs for technical reasons, or because research groups may change affiliation and, consequently, their tools migrate to another address. Another problem is that services may

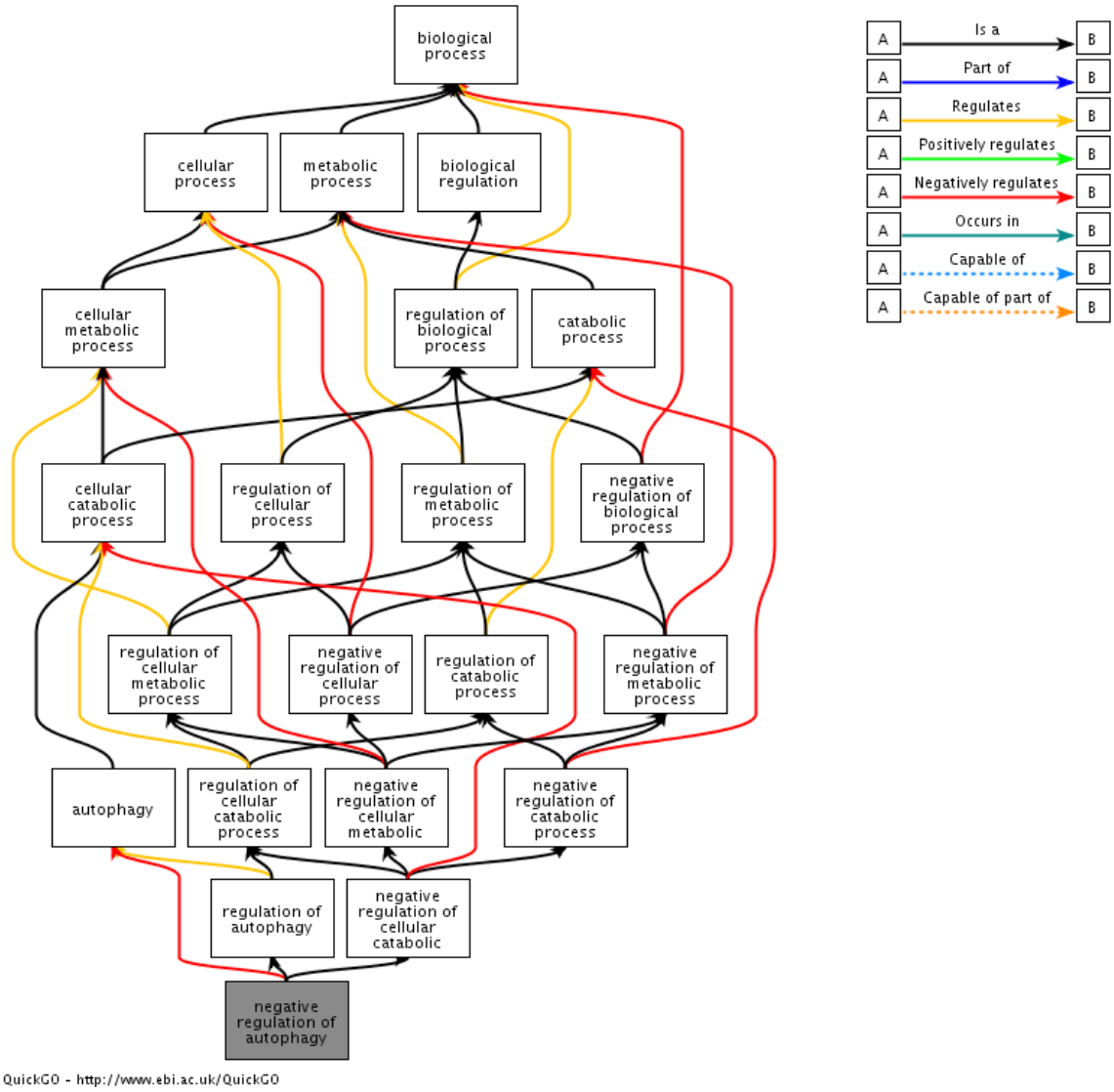

Figure 3: Relations of "outer mitochondrial membrane protein" Gene Ontology term with others. The image is from the QuickGO server [Binns et al., 2009]. Legend explains the possible relations among terms.

be discontinued, after some years of activity, due to the end of the funding to the project or of the doctoral project of the young investigator involved into the development, or for other reasons. The creation of a stable repository of active bioinformatics tools, that may constitute a standard reference for bioinformatics applications, together with a strong training of researchers in the correct use of tools and analysis strategies, represent the next challenge for the standardization of research in omics sciences.

Acknowledgments: A.F. activity has been supported by the Flagship "InterOmics" project (PB.P05) funded by the Italian Ministry of Education, University and Research and Italian National Research Council organizations.

References

[1] Strasser B.J. Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure. 1954–1965. J Hist Biol.. 2010. 43, 623-660.

[2] Brazma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., Stoeckert C., Aach J., Ansorge W., Ball C.A., Causton H.C., et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet., 2001, 29, 365-371.

[3] Taylor C.F., Paton N.W., Lilley K.S., Binz P.A., Julian R.K. Jr, Jones A.R., Zhu W., Apweiler R., Aebersold R., Deutsch E.W., et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol., 2007, 25, 887-893.

[4] Chervitz S.A., Deutsch E.W., Field D., Parkinson H., Quackenbush J., Rocca-Serra P., Sansone S.A., Stoeckert C.J. Jr,

Taylor C.F., Taylor R., Ball C.A. Data standards for Omics data: the basis of data sharing and reuse. Methods Mol Biol., 2011, 719, 31-69.

[5] Baker N.A., Klemm J.D., Harper S.L., Gaheen S., Heiskanen M., Rocca-Serra P., Sansone S.A. Standardizing data. Nat Nanotechnol., 2013, 8, 73-74.

[6] Kim M.S., Pinto S.M., Getnet D., Nirujogi R.S., Manda S.S., Chaerkady R., Madugundu A.K., Kelkar D.S., Isserlin R., Jain S., et al. A draft map of the human proteome. Nature, 2014, 509, 575-581.

[7] Wilhelm M., Schlegl J., Hahne H., Moghaddas Gholami A., Lieberenz M., Savitski M.M., Ziegler E., Butzmann L., Gessulat S., Marx H., et al. Mass-spectrometry-based draft of the human proteome. Nature, 2014, 509, 582-587.

[8] The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res., 2015 (Database issue), 43, D204-D212.

[9] Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res., 2015 (Database issue), 43, D30-D35.

[10] Goodstein D.M., Shu S., Howson R., Neupane R., Hayes R.D., Fazo J., Mitros T., Dirks W., Hellsten U., Putnam N., Rokhsar D.S. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res., 2012 (Database issue), 40, D1178-D1186.

[11] Pagani I., Liolios K., Jansson J., Chen I.M., Smirnova T., Nosrat B., Markowitz V.M., Kyrpides N.C. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res., 2012 (Database issue), 40, D571-D579.

[12] Facchiano A.M., Facchiano A., Facchiano F. Active Sequences Collection (ASC) database: a new tool to assign functions to protein sequences. Nucleic Acids Res., 2003, 31, 379-382.

[13] Wang G., Li X. and Wang Z. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res., 2009 (Database issue), 37, D933-D937.

[14] Fjell C. D., Hancock R.E., Cherkasov, A. AMPer: a database and an automated discovery tool for antimicrobial peptides. Bioinformatics, 2007, 23, 1148-1155.

[15] Di Luca M., Maccari, G., Maisetta G., Batoni G. BaAMPs: the database of biofilm-active antimicrobial peptides Biofouling, 2015, 31, 193-199.

[16] Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard JL, Hix D, Sette A, Peters B. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2015 Jan;43(Database issue), D405-412.

[17] Brusic V, Millot M, Petrovsky N, Gendel SM, Gigonzac O, Stelman SJ. Allergen databases. Allergy, 2003, 58, 1093-1100.

[18] Minkiewicz P., Dziuba J., Iwaniak A., Dziuba M., Darewicz M. BIOPEP database and other programs for processing bioactive peptide sequences. Journal of AOAC International, 2008, 91, 965-980.

[19] Ye J., McGinnis S., Madden T.L. BLAST: improvements for better sequence analysis. Nucleic Acids Res., 2006, 34 (Web Server issue), W6-9.

[20] Pietrokovski S., Henikoff J.G., Henikoff S. The Blocks database--a system for protein classification. Nucleic Acids Res., 1996, 24, 197-200.

[21] Vizcaíno J.A., Côté R.G., Csordas A., Dianes J.A., Fabregat A., Foster J.M., Griss J., Alpi E., Birim M., Contell J., et al. The Proteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res., 2013 (Database issue), 41, D1063-D1069.

[22] Csordas A., Wang R., Rios D., Reisinger F., Foster J.M., Slotta D.J., Vizcaino J.A., Hermjakob H. From Peptidome to PRIDE: Public proteomics data migration at a large scale. Proteomics, 2013, 13, 1692-1695.

[23] Hoogland C., Mostaguir K., Appel R.D., Lisacek F. The World-2DPAGE Constellation to promote and publish gel-based proteomics data through the ExPASy server. J. of Proteomics, 2008, 71, 245-248.

[24] Binns D., Dimmer E., Huntley R., Barrell D., O'Donovan C., Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics, 2009, 25, 3045-3046.

[25] Tanabe M., Kanehisa M. Using the KEGG database resource. Curr Protoc Bioinformatics, 2012, Chapter 1:Unit1.12.

[26] Amberger J.S., Bocchini C.A., Schiettecatte F., Scott A.F., Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res., 2015, 43(Database issue), D789-D798.