Małgorzata Plechawska-Wójcik[1], Magdalena Borys[2]

INTEGRATION MECHANISM OF CHOSEN BIOLOGICAL DATABASES

The paper presents the analysis of several biological database mechanisms. The presented project involves integration of 4 proteomic and genomic databases (EPO-KB, UniProt, NCBI, KEGG). The idea of the work is to build and implement a mechanism which enables fast access of data from those databases. Each database provides different level of access. The presented mechanism is implemented in the application dedicated to Maldi-Tof (Matrix-Assisted Laser Desorption Ionization – Time Of Flight) spectra analysis. It gives access to 4 levels of information about proteins detected in the process of the spectra analysis. The first level is based on m/z values represented by peaks of spectrum. The levels 2-4 deal with the details of proteins, genes and gene paths.

Keywords: *biological databases; mechanism of access; levels of information; proteomic data; genes.*

Малгожата Плехавська-Вуйчик, Магдалена Борис

МЕХАНІЗМ ІНТЕГРАЦІЇ БІОЛОГІЧНИХ БАЗ ДАНИХ

У статті проаналізовано механізми роботи біологічних баз даних. Представлено проект інтеграції 4 протеомних та геномних баз даних – EPO-KB, UniProt, NCBI та KEGG. Така інтеграція є можливою завдяки механізму швидкого доступу до даних цих баз. У кожної з представлених баз є свої рівні доступу. Роботу механізму продемонстровано на прикладі доступу до даних для спектрального аналізу. Механізм надає доступ до 4 рівнів інформації щодо білків, виявлених в процесі спектрального аналізу. Перший рівень базується на пікових значеннях спектру. Рівні 2-4 мають справу з білками, генами та шляхами вивчення генів.

Ключові слова: *біологічні бази даних; механізм доступу; рівні інформації; протеомні дані; гени.*

Рис. 2. Табл. 1. Літ. 16.

Малгожата Плехавська-Вуйчик, Магдалена Борис

МЕХАНИЗМ ИНТЕГРАЦИИ БИОЛОГИЧЕСКИХ БАЗ ДАННЫХ

В статье проанализированы механизмы работы биологических баз данных. Представлен проект интеграции 4 протеомных и геномных баз данных – EPO-KB, UniProt, NCBI и KEGG. Такая интеграция возможна благодаря механизму быстрого доступа к данным этих баз. У каждой из представленных баз существуют свои уровни доступа. Работа механизма продемонстрирована на примере доступа к данным для спектрального анализа. Механизм дает доступ к 4 уровням информации о белках, обнаруженных в процессе спектрального анализа. Первый уровень базируется на пиковых значениях спектра. Уровни 2-4 имеют дело с белками, генами и путями изучения генов.

Ключевые слова: *биологические базы данных; механизм доступа; уровни информации; протеомные данные; гены.*

1. Introduction. Biological databases are a powerful tool enabling easy interpretation of analyzed data. Such databases are maintained and reorganized continuously to keep them up-to-date. Research provides new data and new dependencies con-

[1] Institute of Computer Science, Lublin University of Technology, Poland.
[2] Institute of Computer Science, Lublin University of Technology, Poland.

stantly. Most of databases are equipped with tools supporting the process of data updating and cross references checking. However, the best quality databases are those which hire experts to check references manually. Of course, they are supported by several tools and smaller databases. So far, however, manual checking gives the best results. Bigger databases are composed of several smaller ones. Each contains data dedicated to other type of molecule. They have also dedicated tools.

2. Biological databases. Biological databases are the basic source of information about proteins and their sequences as well as about genes coding them. Those data come from research experiments and their interpretation, publications and other databases. Research centers involved in construction and maintenance of biological databases cooperate with each other and exchange data.

The problem of biological databases is the enormous growth of information and the constant need of standardizing and structuring the stored information. Biological data, in particular on protein, are difficult to manage because of the continuous influx of information and lack of uniform standards of naming and classification. Continuous exchange of data between the databases in different centers allows frequent updates. On the other hand, continuous quality control and consistency of data is required. Also methods automating those processes need to be developed.

Nucleotide records are the basic biological information available in databases. They constitute a source of proteins' sequences contained in databases. Incorrect or incomplete information contained in the nucleotide record may have several consequences [1]:

- Record can be useless if there is a lack of the name of the coding region (CDS). A lack of CDS indicating mRNA segments may prevent from making a scientific discovery;
- If a set of nucleotide sequence features is limited, the appropriate nucleotide record will not contain relevant information about a protein;
- Mistaken description of regions coding the nucleotide sequence containing incorrect information about a protein can be multiplied and spread quickly to other nucleotide and protein databases.

Among the best-known biological databases one can find: UniProt, NBCI, KEGG, EXPASY, HPRD, EPO-KB.

UniProt database (The Universal Protein Resource) [15] is built of 3 parts based on databases: Swiss-Prot, TrEMBL (Translated EMBL) and PIR (Protein Information Resource). The entire UniProt database is a consistent source of information on protein sequences and annotations. It is divided into several parts: UniProt Knowledgebase (UniProtKB), The UniProt Reference Clusters (UniRef), the UniProt Archive (UniParc) and The UniProt Metagenomic and Environmental Sequences (UniMES). UniProtKB database is the central part on the collection containing the data of proteins and their annotations. Uni-Parc (UniProtArchive) database is a non-redundant database containing most of the publicly available protein sequences. UniProtKB database is composed of two sections. The first one covers manually annotated records containing information derived from the literature and subjected to computational analysis carried out by teams of biologists. The second section consists of computer-analyzed records waiting for full manual annotations. UniRef databases contain clustered sequence sets from the UniProt Knowledgebase including alternative splicing, isoforms and selected UniParc records. Fragments of the sequence are grouped in different ways. UniRef100 database combines in one entry identical sequences coming from different organisms. UniRef50 and UniRef90 databases are created by clustering of sequences contained in UniRef100. Each cluster is a sequence containing respectively, at least 50% or 90% sequence compatibility of the longest sequence contained in UniRef.

NCBI (The National Information Center for Bioinformatics) database [9] is a publicly available database, which groups information from 3 different sources: the Swiss-Prot, TrEMBL, RefSeq. The NCBI database is non-redundant and the effort is put to avoid the identical sequences existing.

NCBI is a project whose research includes the organization of genes, sequence analysis, protein domains in order to structure prediction and the human genome map formation. NCBI tools provide also mathematical modeling and analysis of errors derived from inaccurate sequences present in the database. Among the databases, NCBI provides also a set of useful tools. One of them is the Entrez – a tool supporting the search for information about nucleotide and protein sequences (CDD – Conserved Domain Database), protein domains (Domains 3D), protein structures and genome mappings (Genome). It also contains publications on over 600 organisms. Another useful tool is the Map Viewer, which enables users to browse human genome sequences in the form of graphical maps.

EPO-KB database (Empirical Proteomics Ontology Knowledge Base) [3, 7] is the database on protein ontology based on OWL (Web Ontology Language). It contains validated information on biomarkers and their links to protein. It allows identification of proteins regarding their modification. The database operates on UniProt data and is provided with data from the research experiments. EPO-KB enables identification of proteins using M/Z values. The interface enables simultaneous processing of data received from the entire spectrum by specifying the M/Z values. One can also specify the detailed search options such as tolerances, double or triple charges, the type of spectrometry platform and type of biological sample. The database can cope with different combinations of M/Z values pointing to the same protein by using the so-called "biomarkers pools".

KEGG (Kyoto Encyclopedia of Genes and Genomes) [4, 5] knowledge base combines bioinformatics databases containing information about genes, genomes, gene pathways, molecular interactions, proteins, chemical compounds and diseases. This database is used not only to search for information. It also allows modeling and simulation of biological processes.

Among the main elements of the KEGG database one can find: KEGG PATHWAY, KEGG BRITE, KEGG Genes, KEGG LIGAND. KEGG PATHWAY database shows, via diagrams, the molecular interactions and metabolic processes including carbohydrates, amino acids, lipids, phosphates. The database also describes the cellular processes (communication, transportation and travel, growth and death, behavior), information processing and human genetic diseases and drug development. The database is created and updated manually on the basis of published results. KEGG BRITE database presents various aspects of biological systems and contains a hierarchically structured data associated with proteomics, genetics and chemistry. These data are stored in the form of network paths.

3. Proteomic data analysing. Biological databases are often treated as a part of the complex data analysis. For example, it can be useful in the process of mass spectrometry data analysis. Mass spectrometry data need complex, multistep processing. A process of gaining biological information and knowledge from row data is composed of several steps. All those steps need to be performed to get the information which may be helpful in diagnosis or medical treatment tasks.

The comprehensive mass spectra analysis is usually composed of such steps as:

1. Preprocessing [10, 12] to prepare data for main analysis. Depending on the data and used methods, preprocessing can be composed of trimming, binning, denoising, baseline correction, normalization, interpolation. Processes like denoising and baseline correction can be done with local maxima smoothing (Cromwell package [12], PROcess [6], ProteinChip Software [2]), removing some detected peaks (LIMPIC [8]) or moving average with dedicated filters (OpenMS [14]).

2. The core of signal analysis is peak detection. It can be obtained on the basis of local maxima and signal to noise ratio ([2]), area under the peaks curve ([6]), the height (SpecAlign [16]) and the shape of peaks (OpenMS [14]). It is also possible to use functions modeling. An effective peak detection method is based on Gaussian mixture models decomposition performed using expectation-maximization algorithm [13].

3. Classification and choosing the most informative peaks are usually further steps of analysis. The most common classification task is based on supervised learning and it consists of categorizing data into two or more groups. Mass spectrometry data, however, are characterized with high dimensionality. The number of observations is significantly lower than the number of features. That is why researchers usually need to apply dimension reduction techniques [11].

4. Biological interpretation based on dedicated databases is essential part of the analysis, because this part of analysis gives concrete, truly useful results. Results of peaks analysis and classification can be interpreted in this step. It is common to use several databases for single interpretation process. Such approach gives better opportunity to find data and its characteristics.

Many mass spectra analysis tools concentrate mainly on solving mathematical models. The most useful tools, however, are more comprehensive solutions. Such solutions support biologists in data analysis, concerning creating and solving the mass spectrometry data models, choosing the most informative peaks and giving their biological interpretation.

Using proteomic techniques as a way to support early diagnosing of diseases is an opportunity for developing a new way of treatment. There is a group of diseases which needs new treatment and diagnosis approaches. For them, typical ambulatory methods are not always useful. Particularly, this group includes a subgroup of cancer diseases. It is possible to distinguish between ill patients and healthy donors or to check reactions (positive or negative) during medical treatment. It is also possible to look for a stage of diseases progression and looking for significant proteins or genes.

4. Biological interpretation suporting tool

This chapter presents biological interpretation module, which can be used as a separate application or as a part of comprehensive mass spectrometry data analysis tool. Its primary advantage is integration of several large, external protein databases and access to search them in one place. It supports both biological interpretations of data downloaded from outside or transferred directly with the mass spectral analysis tool. The analysis is possible on levels of detail. The levels must be attained in succession and each next level can be achieved after the previous one. The analysis can be canceled at any level. There are also options for supporting the work of a logged user, such as the operation history or saving several different variants of settings.

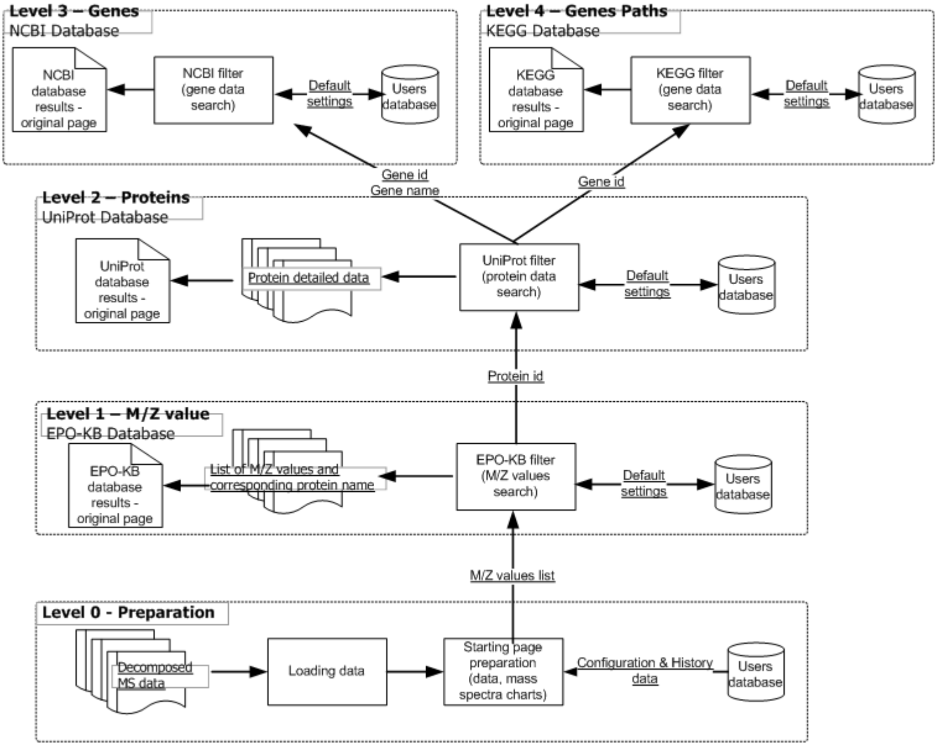
Level 3 – Genes
NCBI Database
NCBI database results - original page
NCBI filter (gene data search)
Default settings
Users database
Level 4 – Genes Paths
KEGG Database
KEGG database results - original page
KEGG filter (gene data search)
Default settings
Users database
Gene id
Gene name
Gene id
Level 2 – Proteins
UniProt Database
UniProt database results - original page
Protein detailed data
UniProt filter (protein data search)
Default settings
Users database
Protein id
Level 1 – M/Z value
EPO-KB Database
EPO-KB database results - original page
List of M/Z values and corresponding protein name
EPO-KB filter (M/Z values search)
Default settings
Users database
M/Z values list
Level 0 - Preparation
Decomposed MS data
Loading data
Starting page preparation (data, mass spectra charts)
Configuration & History data
Users database

Fig. 1. **Functionalities of the application**

Functionalities of the application are presented in Fig.1.:

1. Level 0 is the level responsible for the data loaded into the system. This level enables giving additional search criteria such as the percentage of error tolerance, double or triple charges existence, type of spectrometry platform or type of examined tissue. For registered users there is a possibility to load previously defined settings.

2. Level 1 is defined as the M/Z value level. It is the basic level of the analysis. It allows obtaining the list of proteins and peptides corresponding to M/Z values. The results on this level are obtained from a EPO-KB database [7]. The results are taken directly from the EPO-KB web service. There is also possibility of display the original EPO-KB page with detailed information. This information can include the range of M/Z values, related diseases, the type of genetic material and access to data sources.

3. Level 2, described as the level of proteins, is available after selection of a particular protein/peptide on the level 1. At this leveled detail information about the selected protein/peptide is available. The data are downloaded from the UniProt [15]

web service. The search is done by name and additional criteria, such as species. A user can get such information as a name, the status of the checking process, the name and identifiers of encoding genes, essential features and functional annotations of genes. There is also a link to direct page available to get detailed information contained in the UniProt database.

4. Level 3 is called the level of genes. This level allows displaying the information about the selected genes encoding proteins on the level 2. Presented data are taken from the NCBI database [9]. The results are displayed as a original NCBI page containing information about a specific gene and its role in the body or specific processes, its status and sources. Searching is based on a gene identifier, if it is known or on a gene's name.

5. Level 4 is the level of gene and their paths. It can be achieved directly from the level 2. Details of genes paths are available through the KEGG database [4]. The data are searched on the basis of the ID gene.

5. Results and conclusions. The analyzed dataset contains mass spectrometry data of 5 healthy women. The samples were collected regularly, once a week throughout one menstrual cycle of each woman. The samples were collected one day a week, 3 times during a day (at 8 a.m. – before breakfast, at 10 a.m. – after breakfast, at 2 p.m. – after lunch). Each sample was taken two times and two technical repeats were performed. All 4 repetitions were averaged and next, the preliminary process and decomposition were performed. Such averaging reduces noise and removes outliers. The mean spectrum is presented in Fig. 2.

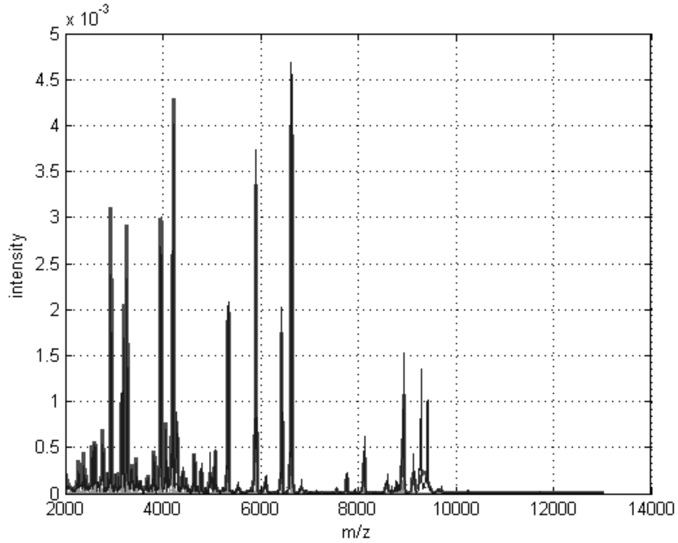

Fig. 2. **The mean spectrum**

Tab. 1 presents the protein results of the first level of the biological analysis. The obtained results were measured with a mean spectrum. Searching was conducted with a tolerance of 0.5%. Single charge data were analyzed.

Table 1. **The biological analysis of the results**

M/Z values	Names of proteins
8916.247	complement c3 frag, vitronectin frag, complement c3, apolipoprotein a-ii
3264	fibrinogen alpha chain frag
8340	complement c3 frag
3955	inter-alpha-trypsin inhibitor heavy chain h4 frag, neurosecretory protein vgf frag
9130	haptoglobin
8790	apolipoprotein c-iii, c-c motif chemokine 13
7767	platelet factor 4
7939	hemoglobin subunit beta
3448	neutrophil defensin 1, neutrophil defensin 3
4474	antithrombin-iii frag
2162	amyloid beta a4 protein
7739	osteopontin frag, platelet factor 4
3159	inter-alpha-trypsin inhibitor heavy chain h4 frag, serum albumin frag
4247	amyloid beta a4 protein
3266	fibrinogen alpha chain frag
3394	neutrophil defensin 3

Mass spectrometry data need for special processing and analyzing. Data specificity makes it hard to analyze and classify. Using application integrating biological databases makes it is possible to gain information about biological context of the analyzed data set. It enables collecting essential information in one place. It also makes the analysis faster and more reliable.

Bibliogaphy:

1. *A.D. Baxevanis and B.F.F. Ouellette.* Bioinformatyka. Podrecznik do analizy genow i bialek. Wydawnictwo Naukowe PWN, 2005.

2. Ciphergen Biosystems: ProteinChip Software Operation Manual, Fremont, CA 94555, 2002.

3. EPO-KB. Empirical proteomics ontology knowledge base. http://www.dbmi.pitt.edu/EPO -KB.

4. Kanehisa Laboratories. Kegg: Kyoto encyclopedia of genes and genomes. http://www.genome.jp/kegg/.

5. *M. Kanehisa, S. Goto, M. Hattori, K. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa.* From genomics to chemical genomics: new developments in Kegg. Nucleic Acids Research, 34:D354-D357, 2006.

6. *Li X., Gentleman R., Lu X., Shi Q., Iglehart J.D., Harris L., Miron A:* SELDI-TOF mass spectrometry protein data, Statistics for Biology and Health, Part I, Bioinformatics and computational biology solutions using R and Bioconductor, 91-109, 2005.

7. *J.Lustgarten, C. Kimmel, H. Ryberg, and W. Hogan.* Epo-kb: a searchable knowledge base of biomarker to protein links. Bioinformatics, 24(11):1418-1419, 2008.

8. *Mantini D., Petrucci F., Pieragostino D., Del Boccio P., Di Nicola M., Di Ilio C., Federici G., Sacchetta P., Comani S., Urbani A.:* LIMPIC: a computational method for the separation of protein signals from noise, BMC Bioinformatics, 2007, 8, 101.

9. NCBI. The national center for bioinformatics information (ncbi) database. http://www.ncbi.nlm.nih.gov/.

10. *Norris J., Cornett D., Mobley J., Anderson M., Seeley E., Chaurand P, Caprioli R.:* Processing MALDI mass spectra to improve mass spectral direct tissue analysis. National institutes of health, USA, 260(2-3): pp. 212-221, 2007.

11. *Plechawska-Wojcik M.:* Biological interpretation of the most informative peaks in the task of mass spectrometry data classification. Studia Informatica. Zeszyty Naukowe Politechniki Slaskiej, seria INFORMATYKA. Vol.32, 2A (96). Wydawnictwo Politechniki Slaskiej, 2011, pp. 213-228.

12. *Plechawska, M., J. Polanska, A. Polanski, M. Pietrowska, R. Tarnawski, P. Widlak, M. Stobiecki, Marczak L.:* Analyze of MALDITOF proteomic spectra with usage of mixture of Gaussian distributions. Man-machine interactions, Advances in Intelligent and Soft Computing. Eds: Cyran, K., Kozielski, S., Peters, J., Stanczyk, U., Wakulicz-Deja, A. Berlin, Springer, 113-120, 2009.

13. *Polanska J., Plechawska M., Pietrowska M., Marczak L.:* Gaussian mixture decomposition in the analysis of MALDI-ToF spectra, Expert Systems, doi: 10.1111/j.1468-0394. 2011. 00582.x, 2011

14. *Sturm M., Bertsch A., Gropl C., Hildebrandt A., Hussong R., Lange E., Pfeifer N., Schulz-Trieglaff O., Zerck A., Reinert K., Kohlbacher O.:* OpenMS – An open-source software framework for mass spectrometry, Bioinformatics, 9, 163, 2008.

15. UniProt Consortium. The universal protein resource (uniprot) database. http://www.uniprot.org/.

16. *Wong J.W., Cagney G., Cartwright H.M.:* SpecAlign – processing and alignment of mass spectra datasets, Bioinformatics, 21, 2088-2090, 2005.

Стаття надійшла до редакції 10.01.2012.