Bioinformatics Integration Framework for Metabolic Pathway Data-Mining

Tomás Arredondo V.[1], Michael Seeger P.[2], Lioubov Dombrovskaia[3],
Jorge Avarias[3] A., Felipe Calderón B.[3], Diego Candel C.[3], Freddy Muñoz R.[3],
Valeria Latorre R.[2], Loreine Agulló[2], Macarena Cordova H.[2], and Luis Gómez[2]

[1] Departamento de Electrónica
e-mail: tarredondo@elo.utfsm.cl
[2] Millennium Nucleus EMBA, Departamento de Química
[3] Departamento de Informática,
Universidad Técnica Federico Santa María
Av. España 1680, Valparaíso, Chile

Abstract. A vast amount of bioinformatics information is continuously being introduced to different databases around the world. Handling the various applications used to study this information present a major data management and analysis challenge to researchers. The present work investigates the problem of integrating heterogeneous applications and databases towards providing a more efficient data-mining environment for bioinformatics research. A framework is proposed and GeXpert, an application using the framework towards metabolic pathway determination is introduced. Some sample implementation results are also presented.

1 Introduction

Modern biotechnology aims to provide powerful and beneficial solutions in diverse areas. Some of the applications of biotechnology include biomedicine, bioremediation, pollution detection, marker assisted selection of crops, pest management, biochemical engineering and many others [3, 14, 22, 21, 29].

Because of the great interest in biotechnology there has been a proliferation of separate and disjoint databases, data formats, algorithms and applications. Some of the many types of databases currently in use include: nucleotide sequences (e.g. Ensemble, Genbank, DDBJ, EMBL), protein sequences (e.g. SWISS-PROT, InterPro, PIR, PRF), enzyme databases (e.g. Enzymes), metabolic pathways (e.g. ERGO, KEGG: Kyoto Encyclopedia of Genes and Genomes database) and literature references (e.g. PubMed) [1, 4, 11, 6, 23]. The growth in the amount of information stored has been exponential: since the 1970s, as of April 2004 there were over 39 billion bases in Entrez NCBI (National Center of Bioinformatics databases), while the number of abstracts in PubMed has been growing by 10,000 abstracts per week since 2002 [3, 13].

The availability of this data has undoubtedly accelerated biotechnological research. However, because these databases were developed independently and are managed autonomously, they are highly heterogeneous, hard to cross-reference, and ill-suited to process mixed queries. Also depending on the database being accessed the data in them is stored in a variety of formats including: a host of graphic formats, RAW sequence data, FASTA, PIR, MSF, CLUSTALW, and other text based formats including XML/HTML. Once the desired data is retrieved from the one of the database(s) it typically has to be manually manipulated and addressed to another database or application to perform a required action such as: database and homology search (e.g. BLAST, Entrez), sequence alignment and gene analysis (e.g. ClustalW, T-Coffee, Jalview, GenomeScan, Dialign, Vector NTI, Artemis) [8].

Beneficial application developments would occur more efficiently if the large amounts of biological feature data could be seamlessly integrated with data from literature, databases and applications for data-mining, visualization and analysis. In this paper we present a framework for bioinformatic literature, database, and application integration. An application based of this framework is shown that supports metabolic pathway research within a single easy to use graphical interface with assisting fuzzy logic decision support. To our knowledge, an integration framework encompassing all these areas has not been attempted before. Early results have shown that the integration framework and application could be useful to bioinformatics researchers.

In Section 2, we describe current metabolic pathway research methodology. In Section 3 we describe existing integrating architectures and applications. Section 4 describes the architecture and GeXpert, an application using this framework is introduced. Finally, some conclusions are drawn and directions of future work are presented.

2 Metabolic Pathway Research

For metabolic pathway reconstruction experts have been traditionally used a time intensive iterative process [30]. As part of this process genes first have to be selected as candidates for encoding an enzyme within a potential metabolic pathway within an organism. Their selection then has to be validated with literature references (e.g. non-hypothetical genes in Genbank) and using bioinformatics tools (e.g. BLAST: Basic Local Alignment Search Tool) for finding orthologous genes in various other organisms. Once a candidate gene has been determined, sequence alignment of the candidate gene with the sequence of the organism under study has to be performed in a different application for gene locations (e.g. Vector NTI, ARTEMIS). Once the genes required have been found in the organism then the metabolic pathway has to be confirmed experimentally in the laboratory. For example, using the genome sequence of *Acidithiobacillus ferrooxidans* diverse metabolic pathways have been determined.

One major group of organisms that is currently undergoing metabolic pathways research is bacteria. Bacteria possess the highest metabolic versatility of the three domains of living organisms. This versatility stems from their expansion into different natural niches, a remarkable degree of physiological and genetic adaptability and their evolutionary diversity. Microorganisms play a main role in the carbon cycle and in the removal of natural and man-made waste chemical compounds from the environment [17]. For example, *Burkholderia xenovorans* LB400 is a bacterium capable of degrading a wide range of PCBs [7, 29].

Because of the complex and noisy nature of the data, any selection of candidate genes as part of metabolic pathways is currently only done by human experts prior to biochemical verification. The lack of integration and standards in database, application and file formats is time consuming and forces researchers to develop *ad hoc* data management processes that could be prone to error. In addition, the possibility of using Softcomputing based pattern detection and analysis techniques (e.g. fuzzy logic) have not been fully explored as an aid to the researcher within such environments [1, 23].

3 Integration Architectures

The trend in the field is towards data integration. Research projects continue to generate large amounts of raw data, and this is annotated and correlated with the data in the public databases. The ability to generate new data continues to outpace the ability to verify them in the laboratory and therefore to exploit them. Validation experiments and the ultimate conversion of data to validated knowledge need expert human involvement with the data and in the laboratory, consuming time and resources. Any effort of data and system integration is an attempt towards reducing the time spent by experts unnecessarily which could be better spent in the lab [5, 9, 20].

The biologist or biochemist not only needs to be an expert in his field as well stay up to date with the latest software tools or even develop his own tools to be able to perform his research [16, 28]. One example of these types of tools is BioJava, which is an open source set of Java components such as parsers, file manipulation tools, sequence translation and proteomic components that allow extensive customization but still require the development of source code [25]. The goal should be integrated user-friendly systems that would greatly facilitate the constructive cycle of computational model building and experimental verification for the systematic analysis of an organism [5, 8, 16, 20, 28, 30]. Sun *et al.* [30] have developed a system, IdentiCS, which combines the identification of coding sequences (CDS) with the reconstruction, comparison and visualization of metabolic networks. IdentiCS uses sequences from public databases to perform a BLAST query to a local database with the genome in question. Functional information from the CDSs is used for metabolic reconstruction. One shortcoming is that the system does not incorporate visualization or ORF (Open Reading Frames) selection and it includes only one application for metabolic sequence reconstruction (BLAST).

Information systems for querying, visualization and analysis must be able to integrate data on a large scale. Visualization is one of the key ways of making the researcher work easier. For example, the spatial representation of the genes within the genome shows location of the gene, reading direction and metabolic function. This information is made available by applications such as Vector NTI and Artemis. The function of the gene is interpreted through its product, normally the protein in a metabolic pathway, which description is available from several databases such as KEGG or ERGO [24]. Usually, the metabolic pathways are represented as a flowchart of several levels of abstraction, which are constructed manually. Recent advances on metabolic network visualization include virtual reality use [26], mathematical model development [27], and a modeling language [19]. These techniques synthesize the discovery of genes and are available in databases such as Expasy, but they should be transparent to the researcher by their seamless integration into the research application environment.

The software engineering challenges and opportunities in integrating and visualizing the data are well documented [16, 28]. There are some applications servers that use a SOAP interface to answer queries. Linking such tools into unified working environment is non-trivial and has not been done to date [2]. One recent approach towards human centered integration is BioUse, a web portal developed by the Human Centered Software Engineering Group at Concordia University. BioUse provided an adaptable interface to NCBI, BLAST and ClustalW, which attempted to shield the novice user from unnecessary complexity. As users became increasingly familiar with the application the portal added shortcuts and personalization features for different users (e.g. pharmacologists, microbiologists) [16]. BioUse was limited in scope as a prototype and its website is no longer available. Current research is continued in the CO-DRIVE project but it is not completed yet [10].

4 Integration Framework and Implementation

This project has focused on the development of a framework using open standards towards integrating heterogeneous databases, web services and applications/tools. This framework has been applied in GeXpert, an application for bioinformatics metabolic pathway reconstruction research in bacteria. In addition this application includes the utilization of fuzzy logic for helping in the selection of the best candidate genes or sequences for specified metabolic pathways. Previously, this categorization was typically done in an *ad hoc* manner by manually combining various criteria such as e-value, identities, gaps, positives and score. Fuzzy logic enables an efficiency enhancement by providing an automated sifting mechanism for a very manually intensive bioinformatics procedure currently being performed by researchers.

GeXpert is used to find, build and edit metabolic pathways (central or peripheral), perform protein searches in NCBI, perform nucleotide comparisons of organisms versus the sequenced one (using tblastn). The application can also perform a search of 3D models associated with a protein or enzyme (using the Cn3D viewer), generation of ORF diagrams for the sequenced genome, generation of reports relating to the advance of the project and aid in the selection of BLAST results using T-S-K (Takagi Sugeno Kang) fuzzy logic.

The architecture is implemented in three layers: presentation, logic and data [18]. As shown in Figure 1, the presentation layer provides for a web based as well as a desktop based interface to perform the tasks previously mentioned.

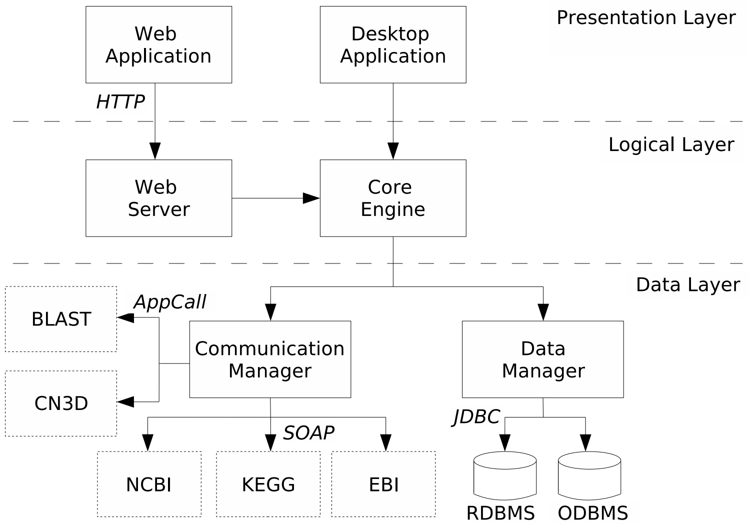


Fig. 1. High Level Framework Architecture

The logical layer consists of a core engine and a web server. The core engine performs bioinformatic processing functions and uses different tools as required: BLAST for protein and sequence alignment, ARTEMIS for analysis of nucleotide sequences, Cn3D for 3D visualization, and a T-S-K fuzzy logic library for candidate sequence selection. The Web Server component provides an HTTP interface into the GeXpert core engine.

As seen in Figure 1, the data layer includes a communications manager and a data manager. The communication manager is charged with obtaining data (protein and nucleotide sequences, metabolic pathways and 3D models) from various databases and application sources using various protocols (application calls, TCP/IP, SOAP). The data manager implements data persistence as well as temporary cache management for all research process related objects.

4.1 Application Implementation: GeXpert

GeXpert [15] is an open source implementation of the framework previously described. It relies on Eclipse [12] builder for multiplatform support.

The GeXpert desktop application as implemented consists of the following:

- Metabolic pathway editor: tasked with editing or creating metabolic pathways, shows them as directed graphs of enzymes and compounds.
- Protein search viewer: is charged with showing the results of protein searches with user specified parameters.
- Nucleotide search viewer: shows nucleotide search results given user specified protein/parameters.
- ORF viewer: is used to visualize surrounding nucleotide ORFs with a map of colored arrows. The colors indicate the ORF status (found, indeterminate, erroneous or ignored).

The GeXpert core component implements the application logic. It consists of the following elements:

- Protein search component: manages the protein search requests and its results.
- Nucleotide search component: manages the nucleotide search requests and its results. In addition, calls the Fuzzy component with the search parameters specified in order to determine the best candidates for inclusion into the metabolic pathway.
- Fuzzy component: attempts to determine the quality of nucleotide search results using fuzzy criteria [1].The following normalized (0 to 1 values) criteria are used: e-value, identities, gaps. Each has five membership functions (very low, low, medium, high, very high), the number of rules used is 243 (3^5).
- ORF component: identifies the ORFs present in requested genome region.
- Genome component: manages the requests for genome regions and its results.

GeXpert communications manager receives requests from the GeXpert core module to obtain and translates data from external sources. This module consists of the following subcomponents:

- BLAST component: calls the BLAST application indicating the protein to be analyzed and the organism data base to be used.
- BLAST parser component: translates BLAST results from XML formats into objects.
- Cn3D component: sends three-dimensional (3D) protein models to the Cn3D application for display.
- KEGG component: obtains and translates metabolic pathways received from KEGG.
- NCBI component: performs 3D protein model and document searches from NCBI.
- EBI component: performs searches on proteins and documentation from the EBI (European Bioinformatic Institute) databases.

GeXpert data manager is tasked with load administrating and storage of application data:

- Cache component: in charge of keeping temporary search results to improve throughput. Also implements aging of cache data.

- Application data component: performs data persistence of metabolic paths, 3D protein models, protein searches, nucleotide searches, user configuration data, project configuration for future usage.

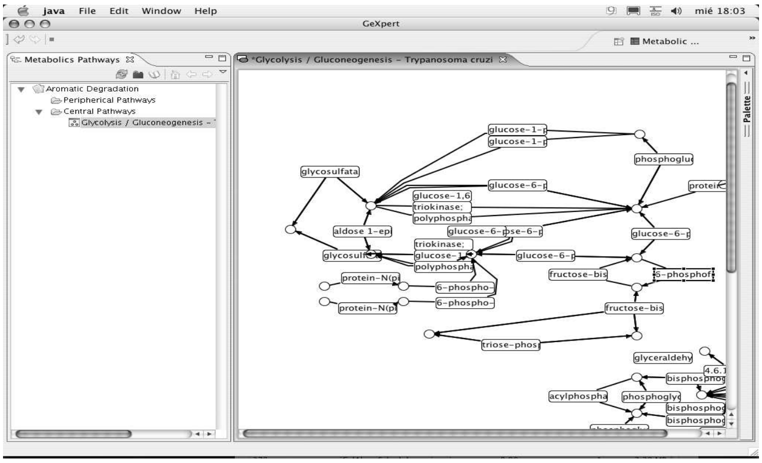
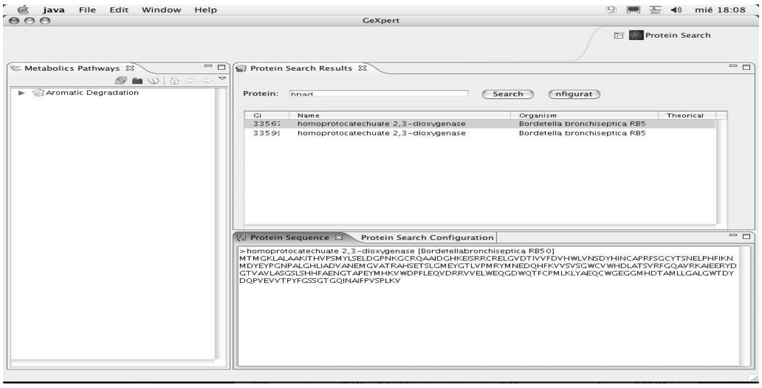


Fig. 2. Metabolic Pathway Viewer

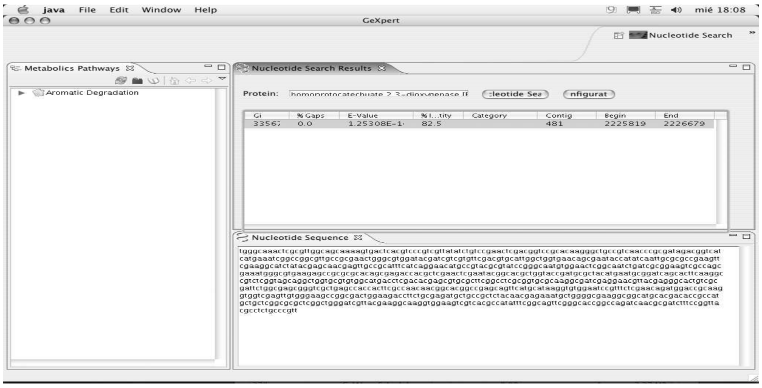


Protein: hnad

Gi	Name	Organism	Theorical
33567	homoprotocatechuate 2,3-dioxygenase	Bordetella bronchiseptica RB5	
3359(	homoprotocatechuate 2,3-dioxygenase	Bordetella bronchiseptica RB5	

```
>homoprotocatechuate 2,3-dioxygenase [Bordetellabronchiseptica RB50]
MTMGKLALAAKITHVPSMYLSELDGPNKGCRQAAIDGHKEISRRCRELGVDTIVVFDVHWLVNSDYHINCAPRFSGCYTSNELPHFIKN
MDYEYPGNPALGHLIADVANEMGVATRAHSETSLGMEYGTLVPMRYMNEDQHFKVVSVSGWCVWHDLATSVRFGQAVRKAIEERYD
GTVAVLASGSLSHHFAENGTAPEYMHKVWDPFLEQVDRRVVELWEQGDWQTFCPMLKLYAEQCWGEGGMHDTAMLLGALGWTDY
DQPVEVVTPYFGSSGTGQINAIFPVSPLKV
```

Fig. 3. Protein Search Viewer



Protein: homoprotocatechuate 2,3-dioxygenase [I

Gi	% Gaps	E-Value	% I...tity	Category	Contig	Begin	End
33567	0.0	1.25308E-1·	82.5		481	2225819	2226679

```
tgggcaaactcgcgttggcagcaaaagtgactcacgtcccgtcgttatatctgtccgaactcgacggtccgcacaagggctgccgtcaacccgcgatagacggtcat
catgaaatcggccggcgttgccgcgaactgggcgtggatacgatcgtcgtgttcgacgtgcattggctggtgaacagcgaataccatatcaattgcgcgccgaagtt
cgaaggcatctatacgagcaacgagttgccgcatttcatcaggaacatgccgtacgcgtatccgggcaatgtggaactcggcaatctgatcgcggaagtcgccagc
gaaatgggcgtgaagagccgcgcgcacagcgagaccacgctcgaactcgaatacggcacgctggtaccgatgcgctacatgaatgcggatcagcacttcaaggc
cgtctcggtagcaggctggtgcgtgtggcatgacctcgacacgagcgtgcgcttcggcctcgcggtgcgcaaggcgatcgaggaacgttacgagggcactgtcgc
gattctggcgagcgggtcgctgagccaccacttcgccaacaacggcacggccgagcagttcatgcataaggtgtggaatccgtttctcgaacagatggaccgcaag
gtggtcgagttgtgggaagccggcgactggaagaccttctgcgagatgctgccgctctacaacgagaaatgctggggcgaaggcggcatgcacgacaccgccat
gctgctcggcgcgctcggctgggatcgttacgaaggcaaggtggaagtcgtcacgccatatttcggcagttcgggcaccggccagatcaacgcgatctttccggtta
cgcctctgcccgtt
```

Fig. 4. Nucleotide Search Viewer

4.2 Application Implementation: GeXpert User Interface and Workflow

GeXpert is to be used in metabolic pathway research work using a research workflow similar to identiCS [30]. The workflow and some sample screenshots are given:

1. User must provide the sequenced genome of the organism to be studied.
2. The metabolic pathway of interest must be created (can be based on the pathway of a similar organism). In the example in Figure 2, the glycolisis/gluconeogenesis metabolic pathway was imported from KEGG.
3. For each metabolic pathway a key enzyme must be chosen in order to start a detailed search. Each enzyme could be composed of one or more proteins (subunits). Figure 3 shows the result of the search for the protein homoprotocatechuate 2,3-dioxygenase.
4. Perform a search for amino acid sequences (proteins) in other organisms. Translate these sequences into nucleotides (using tblastn) and perform an alignment in the genome of the organism under study. If this search does not give positives results it return to the previous step and search another protein sequence. In Figure 4, we show that the protein homoprotocatechuate 2,3-dioxygenase has been found in the organism under study (*Burkholderia xenovorans* LB400) in the chromosome 2 (contig 481) starting in position 2225819 and ending in region 2226679. Also the system shows the nucleotide sequence for this protein.
5. For a DNA sequence found, visualize the ORF map (using the ORF Viewer) and identify if there is an ORF that contains a large part of said sequence. If this is not the case go back to the previous step and chose another sequence.
6. Verify if the DNA sequence for the ORF found corresponds to the chosen sequence or a similar one (using blastx). If not choose another sequence.
7. Establish as found the ORF of the enzyme subunit and start finding in the surrounding ORFs sequences capable of coding the other subunits of the enzyme or other enzymes of the metabolic pathway.

8. For the genes that were not found in the surrounding ORFs repeat the entire process.
9. For the enzyme, the 3D protein model can be obtained and viewed if it exists (using CN3D as integrated into the GeXpert interface).
10. The process is concluded by the generation of reports with the genes found and with associated related documentation that supports the information about the proteins utilized in the metabolic pathway.

5 Conclusions

The integrated framework approach presented in this paper is an attempt to enhance the efficiency and capability of bioinformatics researchers. GeXpert is a demonstration that the integration framework can be used to implement useful bioinformatics applications. GeXpert has so far shown to be a useful tool for our researchers; it is currently being enhanced for functionality and usability improvements. The current objective of the research group is to use GeXpert in order to discover new metabolic pathways for bacteria [29].

In addition to our short term goals, the following development items are planned for the future: using fuzzy logic in batch mode, web services and a web client, blastx for improved verification of the ORFs determined, multi-user mode to enable multiple users and groups with potentially different roles to share in a common research effort, peer to peer communication to enable the interchange of documents (archives, search results, research items) thus enabling a network of collaboration in different or shared projects, and the use of intelligent/evolutionary algorithms to enable learning based on researcher feedback into GeXpert.

Acknowledgements

This research was partially funded by Fundación Andes.

References

1. Arredondo, T., Neelakanta, P.S., DeGroff, D.: Fuzzy Attributes of a DNA Complex: Development of a Fuzzy Inference Engine for Codon-'Junk' Codon Delineation. Artif. Intell. Med. **35** 1-2 (2005) 87-105
2. Barker, J., Thornton, J.: Software Engineering Challenges in bioinformatics. Proceedings of the 26th International Conference on Software Engineering, IEEE (2004)
3. Bernardi, M., Lapi, M., Leo, P., Loglisci, C.: Mining Generalized Association Rules on Biomedical Literature. In: Moonis, A. Esposito, F. (eds): Innovations in Applied Artificial Intelligence. Lect. Notes Artif. Int. **3353** (2005) 500-509
4. Brown, T.A.: Genomes. John Wiley and Sons, NY (1999)
5. Cary M.P., Bader G.D., Sander C.: Pathway information for systems biology (Review Article). FEBS Lett. **579** (2005) 1815-1820,
6. Claverlie, J.M.: Bioinformatics for Dummies. Wiley Publishing (2003)

7. Cámara, B., Herrera, C., González, M., Couve, E., Hofer, B., Seeger, M.: From PCBs to highly toxic metabolites by the biphenyl pathway. Environ. Microbiol. (6) (2004) 842-850
8. Cohen, J.: Computer Science and Bioinformatics. Commun. ACM **48** (3) (2005) 72-79
9. Costa, M., Collins, R., Anterola, A., Cochrane, F., Davin, L., Lewis, N.: An in silico assessment of gene function and organization of the phenylpropanoid pathway metabolic networks in *Arabidopsis thaliana* and limitations thereof. Phytochem. **64** (2003) 1097-1112
10. CO-Drive project: http://hci.cs.concordia.ca/www/hcse/projects/CO-DRIVE/
11. Durbin, R.: Biological Sequence Analysis, Cambridge, UK (2001)
12. Eclipse project: http://www.eclipse.org
13. Entrez NCBI Database: www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
14. Gardner, D.,: Using genomics to help predict drug interactions. J Biomed. Inform. **37** (2004) 139-146
15. GeXpert sourceforge page: http://sourceforge.net/projects/gexpert
16. Javahery, H., Seffah, A., Radhakrishnan, T.: Beyond Power: Making Bioinformatics Tools User-Centered. Commun. ACM **47** 11 (2004)
17. Jimenez, J. I., Miambres, B., Garca, J., Daz, E.: Genomic insights in the metabolism of aromatics compounds in Pseudomonas. In: Ramos, J. L. (ed): Pseudomonas, vol. 3. NY: Kluwer Academic Publishers, (2004) 425-462
18. Larman, C.: Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and Iterative Development. Prentice Hall PTR (2004)
19. Loew, L. W., Schaff, J. C.: The Virtual Cell: a software environment for computational cell biology. Trends Biotechnol. **19** 10 (2001)
20. Ma H., Zeng A., Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics **19** (2003) 270-277
21. Magalhaes, J., Toussaint, O.: How bioinformatics can help reverse engineer human aging. Aging Res. Rev. **3** (2004) 125-141
22. Molidor, R., Sturn, A., Maurer, M., Trajanosk, Z.: New trends in bioinformatics: from genome sequence to personalized medicine. Exp. Gerontol. **38** (2003) 1031-1036
23. Neelakanta, P.S., Arredondo, T., Pandya, S., DeGroff, D.: Heuristics of AI-Based Search Engines for Massive Bioinformatic Data-Mining: An Example of Codon/Noncodon Delineation Search in a Binary DNA Sequence, Proceeding of IICAI (2003)
24. Papin, J.A., Price, N.D., Wiback, S.J., Fell, D.A., Palsson, B.O.: Metabolic Pathways in the Post-genome Era. Trends Biochem. Sci. **18** 5 (2003)
25. Pocock, M., Down, T., Hubbard, T.: BioJava: Open Source Components for Bioinformatics. ACM SIGBIO Newsletter **20** 2 (2000) 10-12
26. Rojdestvenski, I.: VRML metabolic network visualizer. Comp. Bio. Med. **33** (2003)
27. SBML: Systems Biology Markup Language. http://sbml.org/index.psp
28. Segal, T., Barnard, R.: Let the shoemaker make the shoes - An abstraction layer is needed between bioinformatics analysis, tools, data, and equipment: An agenda for the next 5 years. First Asia-Pacific Bioinformatics Conference, Australia (2003)
29. Seeger, M., Timmis, K. N., Hofer, B.: Bacterial pathways for degradation of polychlorinated biphenyls. Mar. Chem. **58** (1997) 327-333
30. Sun, J., Zeng, A.: IdentiCS - Identification of coding sequence and in silico reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. BMC Bioinformatics **5:112** (2004)