Combination of a data warehouse concept with web services for the establishment of the *Pseudomonas* systems biology database SYSTOMONAS

Claudia Choi[1], Richard Münch[1], Boyke Bunk[1], Jens Barthelmes[2], Christian Ebeling[2], Dietmar Schomburg[2], Max Schobert[1], Dieter Jahn[1*]

[1]Institut für Mikrobiologie, Technische Universität Braunschweig, Spielmannstraße 7, D-38106 Braunschweig, Germany

[2]Institut für Biochemie, Universität zu Köln, Zülpicher Straße 47, D-50674 Köln, Germany

Summary

Systems biology requires the integration of data from various sources and their combined interpretation using different bioinformatics tools. Integration of different biological databases, however, is often problematic due to their semantic and structural diversity. Moreover, necessary continuous updates of both the structure and content of a database provide further challenges for an integration process. We established the novel database SYSTOMONAS for SYSTems biology of pseudOMONAS by integrating heterogeneous data from highly different external resources including BioCyc, BRENDA, ENZYME, Pseudomonas Genome Database v2, KEGG, and PRODORIC. For this purpose we combined a data warehouse concept with the advantages of web services. This hybrid approach benefits from the fast performance and data consistency provided by the data warehouse system and from the up-to-dateness ensured by use of dynamic web services. The data warehouse part is realized by ETL processes (Extract, Transform, Load), during which data are checked for consistency and standardized to ensure their integrity. While accessing SYSTOMONAS via the internet, parts of the data warehouse content are dynamically enriched using the web service part of the system via SOAP (originally for Simple Object Access Protocol) interfaces with BRENDA, KEGG and PRODORIC. SYSTOMONAS is designed to integrate in-house experimental high-throughput data with up-to-date information available in the mentioned public databases. SYSTOMONAS also serves as a repository for the prediction of metabolic and regulatory networks. SYSTOMONAS is accessible at `http://www.systomonas.de`.

1 Introduction

1.1 Database Integration

During the last decade results from genome sequencing projects and high-throughput transcriptomics, proteomics, and metabolomics studies lead to an explosion of biological data. Moreover, the systematic annotation of gene and protein functions from the literature required data storage and management systems. As a consequence a whole variety of novel databases with different structures and functions were evolving. However, current scientific challenges force

*Corresponding author: d.jahn@tu-bs.de

researchers to consult various of these repositories and analysis tools in a complex and time-consuming process. Therefore, systems are desired, that integrate different data resources and tools under one roof. Such systems should be readily accessible via the internet and minimize the efforts of the user. The diversity and heterogeneity of biological data on one hand and the inconsistency of biological databases on the other hand hamper the integration process. Moreover, keeping the information content and the database structure up-to-date are the major challenges for the implementation of an integrative database system [1].
Some integrated database systems have recently been established such as MaGe [2], BioCyc [3], MicrobesOnline [4], MBGD [5]. In contrast to that BioMOBY [6], and myGrid [7] provide data retrieval via different web services. The two basic alternative technologies realized with these data resources are data warehouse systems and web service applications [1].

1.2 The Data Warehouse Concept

The major function of a data warehouse is to load and translate data from different external sources into one large database. For this process imported data are carefully mapped using ETL-processes (Extract, Transform, Load). Thereafter, all information of the data warehouse is quickly accessible.
In spite of this clear concept practical realization often meets technical difficulties. Setting up the initial unified data model requires careful consideration of the handled data and the required integration abilities. In the second step appropriate software is required for a successful transfer of external data to the unified data model. Since heterogeneous data are integrated into one unified system, usually software development is necessary. Further advantages of the data warehouse concept are the storage of metadata and newly deduced data such as new enzyme annotations derived from comparative genomics studies. Unfortunately, re-imports of external data or continuous updates are required for keeping the data up-to-date. Furthermore, the data extraction machinery needs to be adopted to every structural change of the desired data.
The concept of data warehousing is already realized for several integrative databases including MaGe [2], MicrobesOnline [4], MBGD [5], and Atlas [8].

1.3 Web services

In contrast to the outlined data warehouse, web services dynamically retrieve data of interest from various up-to-date resources. Bioinformatics related web services are provided by several established research institutes (EBI [9], NCBI [10]). They provide useful tools (e. g. MUSCLE [11]; BLAST [12]), and databases (e. g. KEGG [13], BRENDA [14], PRODORIC [15]). Web services are already used for the integration of some biological data resources. BioMOBY [6] and MyGrid [7] are examples of such assembled systems.

In general, a web service is characterized by a well-defined APplication Interface (API) and a Uniform Resource Identifier (URI), which can be registered at the Universal Description, Discovery and Integration (UDDI) registry (see Figure 1a). The underlying methods provided by the web service are well-defined in a WSDL (Web Service Description Language) document. Hereby, the format and type of values for the request and the response are specified as simple data types or even as whole data objects. The currently used binding styles for WSDL are

RPC/encoded and document/literal. For reasons of Web Service Interoperability (WS-I) Basic Profile compliance the document/literal is recommended [16]. Among several transmission standards, that ensure correct data exchange between web service providers and users, SOAP (originally for Simple Object Access Protocol) is one important data communication protocol. It is based on the eXtended Markup Language (XML). SOAP 1.2 consists mainly of three parts: SOAP-envelope, SOAP-header and SOAP-body, the latter containing the real data (see Figure 1b). Communication is based on a server/client model. Requests of the SOAP-client are responded by the SOAP-server.
There are several advantages for retrieving data using SOAP. First of all, if data are changed at the remote database site, the new data can be instantaneously transmitted (*ad hoc* service). Secondly, due to its XML based protocol the transfer is platform and programming language independent. Thirdly, all changes in the database structure of the remote resource need to be in accordance with the defined API. This way structural changes of the remote data repository are avoided or performed on the server site. However, the feasibility of the approach with regard to time and quality is highly dependent on the quality of the implementation of the service and the technical standard of the employed remote systems. Considering the evolution of hardware, network, and software techniques this last point will be of minor relevance in future. Nevertheless, both techniques have important advantages which are complementary.

In this work, we have focussed on data integration for Gram negative proteobacteria of the *Pseudomonas* clade. A database system combining both techniques, data warehousing and web services, was implemented, thereby providing a comprehensive and up-to-date system. To our knowledge this hybrid approach for biological data resources is new in this extend. The systems was named SYSTOMONAS for SYSTems biology of pseudOMONAS. Our goals were (i) the inclusion of available high-throughput and genome data, (ii) the integration of various data sources using the outlined combinatorial approach, (iii) comparative genomics studies, (iv) the reconstruction of metabolic networks, and (v) a user-friendly access to these data via internet.

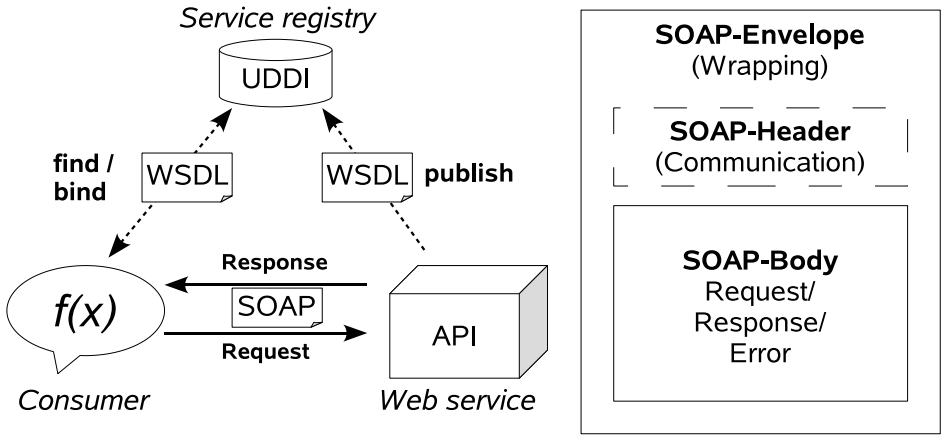

Figure 1: (a) Schema of web services in general. Web services are characterized by a well-defined application interface (API), a web service definition language (WSDL) document, in which methods are explicitly specified, a communication protocol (e. g. SOAP) between consumer and web service provider, and an identifier (uniform resource identifier: URI), which can be registered at the universal description, discovery and integration (UDDI) registry. (b) SOAP protocol structure. The SOAP protocol is composed of the optional SOAP-Header holdig information about the communication process, the SOAP-Body, which contains the 'real' data, and the SOAP-Envelope, which encases these two elements.

2 The Integrative Database SYSTOMONAS

2.1 Challenges During Implementation

Ideally, an information resource for the biological researcher should be easily and quickly accessible via a user-friendly interface. In addition, it should contain as much actual and accurate data as possible. Moreover, information access should not require the consultation of several databases and application tools. As outlined above, data integration is required to meet these claims.
Probably one of the most discussed subjects in the context of diversity and heterogeneity of biological data is the nomenclature of genes. For example, the gene 'hemA' is also found designated as 'hem1', another completely different name is 'glutr'. Dashes, dots, spaces might be inserted. Upper and lower cases might be neglected. The naming in biology remains a creative process. In order to ensure that different people address the very same gene, identifiers for genes were introduced. The so-called locus_ids can be attached to a gene as a unique symbol for a specific gene and its corresponding protein, i. e. the 'PA' numbers for genes in *Pseudomonas aeruginosa*. Unfortunately, genes are not always referenced to their unique identifiers. Therefore, mapping genes from different data resources requires careful considerations. These mapping procedures are easily performed in a data warehouse system.
A general problem of data storage and subsequent data integration is the inconsistency and false-prediction of employed data. Manual curation might be one of the error sources. For example, a wrong DNA-sequence might be given, an inappropriate name might be assigned or a double entry might be inserted. Thorough and accurate controlling is needed to expose the ambiguities and inconsistencies. This maintenance process requires expertise and should be performed foremost in the original database. By using web services it is possible to immediately utilize recently corrected information from the original databases.

2.2 Database architecture

To establish a database covering the knowledge on the molecular biology of the medically and ecologically relevant group of Gram negative bacteria, the pseudomonads, we designed a database model, which combines a data warehouse system with web services. Basic information on genes and proteins, metabolic reactions and pathways, enzyme annotations, metabolome and proteome data along with experimental details were integrated into the core data warehouse. Most of these data are not susceptible to frequent changes. All further currently fast evolving data are retrieved on-the-fly via SOAP including transcription factor binding sites, enzyme features and disease information (see Figure 2).

2.2.1 Data warehouse part

The data warehouse of SYSTOMONAS consisted of flat files from external databases, a relational database system and software for extracting, mapping, and loading the data to the database SYSTOMONAS (see Figure 2). In order to extract the external data in different file formats more readily, all data were stored in the intermediate data container metabold. For retrieving metabolic network information the ligand database of KEGG [13], the ENZYME

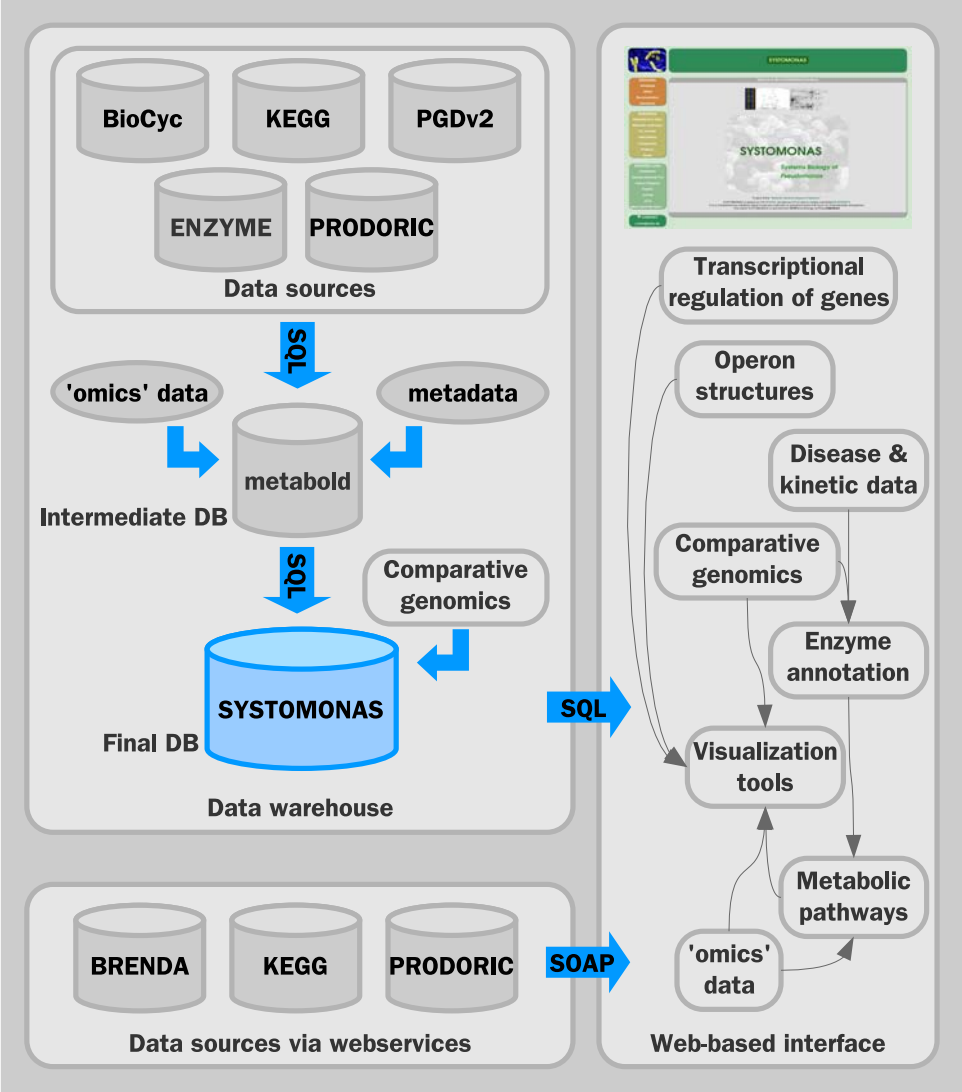

Figure 2: Architecture of the SYSTOMONAS system. Two different techniques - data warehousing and web services - complement in order to improve integration of heterogeneous data and to retrieve up-to-date information.

database [17], BioCyc ,[3] and BRENDA [14] were parsed and stored temporarily in metabold. Up to this step imported data were not restricted to pseudomonads. Genomic information on pseudomonads were retrieved from PRODORIC [15], the Pseudomonas Genome Database v2 (PGDv2) [18]. In-house experimental metabolome data were also added into metabold. Corresponding entries were mapped, double entries were avoided, complementing data were amended. After integrating the data in metabold all data specifically on pseudomonads were imported into SYSTOMONAS. The software BLAST [12] and stretcher (EMBOSS [19]) were applied to identify homologous gene pairs. The enzyme annotation of the gene products were transferred to each homologous gene product partner. The additional tool metaSHARK [20] provided further enzyme annotation, which was integrated as outlined above. Results of this comparative genomic studies were deposited in SYSTOMONAS [21].

2.2.2 Web service part

For SYSTOMONAS appropriate SOAP service clients of PRODORIC, BRENDA, and KEGG were implemented (see Table 1). PRODORIC provides operon structures, relations between transcription factors and their target genes, and results from differential gene expression experiments. BRENDA provides information on enzymes inlcuding kinetic data and relations to diseases. KEGG provides pathway maps with user-defined colouring. The dynamic transfer of external data are realized on the website of SYSTOMONAS.

Table 1: External web services implemented in SYSTOMONAS via SOAP. Eight distinct web services from three different SOAP servers (BRENDA, PRODORIC, KEGG) are included in order to complement the data warehouse system.

Database	BRENDA	PRODORIC	KEGG
Documentation	www.brenda.uni-koeln.de/soap	www.prodoric.de/soap	www.genome.jp/kegg/soap
WSDL site	http://134.95.151.171/soap/brenda.wsdl	http://134.169.104.13/webservice/prodoric.wsdl	http://soap.genome.jp/KEGG.wsdl
Desired data, functions	1. Kinetics 2. Literature 3. Diseases	4. Operons 5. Transcription factors, DNA binding sites 6. Experimental conditions for expression profile experiments 7. Expression profiles experiments	8. Visualization of metabolic pathway maps
Methods	1. getKmValue() 2. getReferenceById() 3. getDisease()	4. getOperon() 5. getRegulatorsFromGene() 6. getProfile() 7. getProfileParameter()	8. get_html_of_marked_athway_by_objects()
Input parameters	1. ligandStructureId:string; organism:string; ecNumber:string 2. id:string 3. ecNumberxsd:string	4. params:OperonParams 5. params:RegulatorParams 6. params:ProfileAcc 7. params:ProfileParameterParams	8. pathway_id:string; object_id_list:ArrayOfstring; fg_color_list:ArrayOfstring; fg_color_list:ArrayOfstring; bg_color_list:ArrayOfstring
Output parameters	1. return:kmValue 2. return:elementaryReference 3. return:disease	4. return:ArrayOfOperonResult 5. return:ArrayOfRegulatorResult 6. return:ArrayOfProfileResult 7. return: ArrayOfProfileParameterResult	8. return:string

Web site generation was done with the scripting language PHP version 5.1 configured with the SOAP extension (`www.php.net`). The employed PHP served as a proficient language for multiple tasks. Beside its function in webpage creation PHP provided an interface to SOAP and for our database management system PostgreSQL (`www.postgresql.org`) for extracting, transforming and loading the data for the database system. In connection to the SOAP interface PHP is the preferred programming language over Java, since less prerequisites are required. Firstly, while using Java several libraries (e. g. Apache AXIS libraries, XML parser) are necessary for the employment of SOAP at the client's site. Secondly, client proxy classes have to be generated with the WSDL2Java tool in a next step. Finally, codes for the call of a web service in Java and PHP are comparable in size. An example code is given in Listing 1 and its implementation at the website (see Figure 3). Unfortunately, the generation of a WSDL-document for SOAP servers has currently to be done manually with PHP, whereas for Java the most commonly used Apache Tomcat server automatically creates this document. For setting up a SOAP server in Java such a functional servlet container has to be used. However, PHP offers several functions to support SOAP since version 5.1 on the server's site. Due to its generic implementation, this integration model can be applied to other integrated databases on organism groups without major efforts.

Listing 1: Example code of the SOAP client in PHP (a) and JAVA (b) including the output (c). The size of the programming code of both languages is similar. The website of this web service is shown in Figure 3.

(a) PHP

```
function soap_gene_reg($locus_id) {
      $client = "http://134.169.104.13/webservice/prodoric.wsdl";
      $return =
      $client->getRegulatorsFromGene((object) array('idtype' =>"orf", 'id' => $locus_id));
      if (sizeof($return->RegulatorResult)==1)
      $return->RegulatorResult=array($return->RegulatorResult);
      return $return;
}

$return = soap_gene_reg("PA4666"):
for ($i=0; $i<sizeof($return->RegulatorResult); $i++) {
      print  ($return->RegulatorResult[$i]->protein_acc . ",\t\t".
            $return->RegulatorResult[$i]->short_name . ",\t\t".
            $return->RegulatorResult[$i]->sequence . ",\t\t".
            $return->RegulatorResult[$i]->references . ";\n");
}
```

(b) Java

```
/* First proxy classes were generated by WSDL2Java
  * e. g. Prodoric_webserviceLocator, ProdoricPortType,
  *        ParamsType, RegulatorResultType.
  */

Prodoric_webservice ws = new Prodoric_webserviceLocator();
ParamsType pt = new ParamsType();
pt.setIdtype("orf");
pt.setId("PA4666");
ProdoricPortType prod = ws.getprodoricPort();

for (RegulatorResultType rrt: prod.getRegulatorsFromGene(pt)){
   System.out.print(rrt.getProtein_acc() + ",\t\t" +
                        rrt.getShort_name() + ",\t\t" +
                        rrt.getSequence() + ",\t\t" +
                        rrt.getReferences() + ";\n");
}
```

(c) Output for both codes:

```
PR00175413, Anr,  TTGAA,            12073043;
PR00175413, Anr,  TTGTT,            12073043;
PR00177747, NarL, TGTTCAT,          12073043;
PR00177747, NarL, TGTCTAT,          12073043;
PR00344205, IHF,  CAAGGGATTTGTTCA,  12073043;
PR00344205, IHF,  CAATACATCGGCAAT,  12073043;
PR00344205, IHF,  CAATGGTTGTCCTGC,  12073043;
```

2.2.3 The Website of SYSTOMONAS

One of our objectives was to provide a user-friendly access to the integrated *Pseudomonas* data. This was realized by providing SYSTOMONAS access at the website `www.systomonas.de`. Without further installation the biologist can readily browse through 'omics' data, integrated data of heterogeneous data sources, comparative genomics data, and reconstructed metabolic networks (see Figure 3). Several tools are provided, to visualize metabolic networks, to analyse metabolome data, to explore gene structures, to align orthologous protein sequences, or to predict further transcription factor binding sites. When necessary web services are dynamically implemented at gene, protein, pathway, and interaction entries (see Figure 3).
For providing the website the free webserver Apache 2.0 was installed (http://apache.org). In

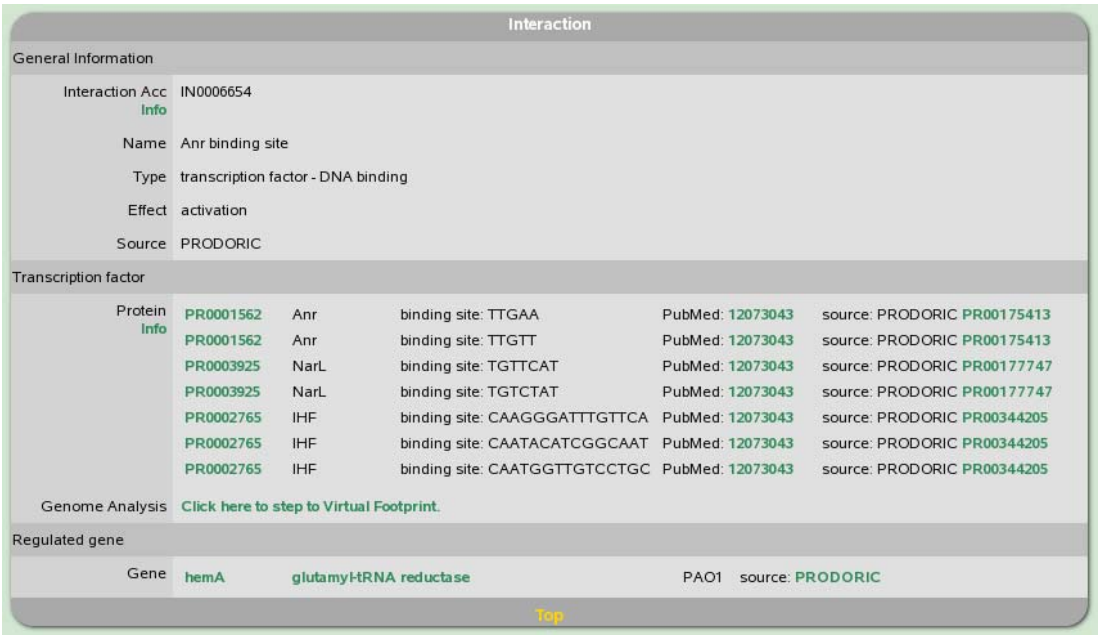

Figure 3: An example for implemented web services at the SYSTOMONAS website. Screenshot of the transcriptional regulation of the *Pseudomonas aeruginosa* gene 'hemA'. PRODORIC protein accession numbers of the transcription factors, transcription factor binding sites, and PubMed references are retrieved via SOAP.

addition to the SOAP implementation and scripts for the data integration in the data warehouse system, website generation was performed with PHP version 5.1 configured with the SOAP extension (`www.php.net`). The data warehouse was organized within the open source object-relational database management system PostgreSQL (`www.postgresql.org`).

3 Comparison of SYSTOMONAS to other database systems

In contrast to integrative database systems provided by MaGe [2], BioCyc [3], MicrobesOnline [4], and MBGD [5] SYSTOMONAS combines several freely accessible databases and tools for the enhanced analysis and understanding of eight genome sequenced *Pseudomonas* species and strains. Firstly, one advantage is the improved annotation of enzymes via the combination of information from KEGG, PRODORIC, PGDv2, BRENDA, ENZYME and BioCyc. Secondly, SYSTOMONAS provides several user-friendly tools from other databases such as Virtual Footprint (PRODORIC) and Genome Explorer (BRENDA). These tools are useful for transcriptional network prediction and visualization. Thirdly, KEGG maps are readily accessible from SYSTOMONAS via SOAP, which is not possible using BioCyc. Fourthly, since we focus on comparative studies, we provide the multiple alignment tool MUSCLE [11] in SYSTOMONAS as well as the alignment viewer Jalview [22]. In combination, both tools are capable of elaborated multiple protein sequence alignments and the generation of phylogenetic trees. Furthermore, our own visualization tool PathCompare shows comparative metabolic KEGG-pathways including all annotated enzymes for eight *Pseudomonas* species at the same time. Fifthly, high-throughput data along with an appropriate analysis tool as well as gene-regulatory data are included. In summary, the main purpose of SYSTOMONAS is to combine information and tools in order to facilitate data access and analysis for all pseudomonads with completed genome determination.

4 Conclusion

- SYSTOMONAS provides an integrated platform of important databases and bioinformatic tools for the systems biology analysis of a medically and ecologically important group of bacteria, the pseudomonads.
- The up-to-date content and fast performance are achieved by a novel combination of a data warehouse concept with web services mediated by SOAP.

References

[1] LD Stein. Integrating biological databases. *Nat. Rev. Genet.*, 4:337–345, 2003.

[2] D Vallenet, L Labarre, Z Rouy, V Barbe, S Bocs, S Cruveiller, A Lajus, G Pascal, C Scarpelli, and C Mdigue. MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res*, 34(1):53–65, 2006.

[3] R Caspi, H Foerster, CA Fulcher, R Hopkinson, J Ingraham, P Kaipa, M Krummenacker, S Paley, J Pick, SY Rhee, C Tissier, P Zhang, and PD Karp. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, 34(Database issue):D511–D516, 2006.

[4] EJ Alm, KH Huang, MN Price, RP Koche, K Keller, IL Dubchak, and AP Arkin. The MicrobesOnline Web site for comparative genomics. *Genome Res*, 15(7):1015–1022, Jul 2005.

[5] I Uchiyama. MBGD: microbial genome database for comparative analysis. *Nucleic Acids Res*, 31(1):58–62, Jan 2003.

[6] M Wilkinson and M Links. BioMOBY: an open source biological web services proposal. *Brief Bioinform*, 3(4):331–41, Dec 2002.

[7] RD Stevens, AJ Robinson, and CA Goble. myGrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19 Suppl 1:i302–4, 2003.

[8] SP Shah, Y Huang, T Xu, MMS Yuen, J Ling, and BFF Ouellette. Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics*, 6:34, 2005.

[9] S Pillai, V Silventoinen, K Kallio, M Senger, S Sobhany, J Tate, S Velankar, A Golovin, K Henrick, P Rice, P Stoehr, and R Lopez. SOAP-based services provided by the European Bioinformatics Institute. *Nucleic Acids Res.*, 33:W25–W28, 2005.

[10] http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html.

[11] RC Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32:1792–1797, 2004.

[12] S McGinnis and TL Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, 32:W20–W25, 2004.

[13] M Kanehisa, S Goto, M Hattori, KF Aoki-Kinoshita, M Itoh, S Kawashima, T Katayama, M Araki, and M Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34:D354–D357, 2006.

[14] I Schomburg, A Chang, C Ebeling, M Gremse, C Heldt, G Huhn, and D Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, 32:D431–D433, 2004.

[15] R Münch, K Hiller, H Barg, D Heldt, S Linz, E Wingender, and D Jahn. PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res.*, 31:266–269, 2003.

[16] PBT Neerincx and JAM Leunissen. Evolution of web services in bioinformatics. *Brief Bioinform*, 6(2):178–188, 2005.

[17] A Bairoch. The ENZYME database in 2000. *Nucleic Acids Res.*, 28:304–305, 2000.

[18] GL Winsor, R Lo, SJH Sui, KSE Ung, S Huang, D Cheng, WKH Ching, REW Hancock, and FSL Brinkman. Pseudomonas aeruginosa Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation. *Nucleic Acids Res.*, 33:D338–D343, 2005.

[19] P Rice, I Longden, and A Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16:276–277, 2000.

[20] JW Pinney, MW Shirley, GA McConkey, and DR Westhead. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. *Nucleic Acids Res.*, 33:1399–1409, 2005.

[21] C Choi, R Münch, S Leupold, J Klein, I Siegel, B Thielen, B Benkert, M Kucklick, M Schobert, J Barthelmes, C Ebeling, I Haddad, M Scheer, A Grote, K Hiller, B Bunk, K Schreiber, I Retter, D Schomburg, and D Jahn. SYSTOMONAS - an integrated database for systems biology analysis of pseudomonas. *Nucleic Acids Res*, in press, 2007.

[22] M Clamp, J Cuff, SM Searle, and GJ Barton. The Jalview Java alignment editor. *Bioinformatics*, 20:426–427, 2004.