InSilico Proteomics System: Integration and Application of Protein and Protein-Protein Interaction Data using Microsoft .NET

Wolfgang Straßer[1], Doris Siegl[1], Kamil Önder[2] and Johann Bauer[2]

[1]Upper Austrian University of Applied Sciences,
Research Center Hagenberg,
Hauptstraße 117, 4232 Hagenberg, Austria.

[2]Paracelsus Medical Private University,
Department of Dermatology,
Müllner Hauptstraße 48, Salzburg, Austria.

Summary

In the last decades, biological databases became the major knowledge resource for researchers in the field of molecular biology. The distribution of information among these databases is one of the major problems. An overview about the subject area of data access and representation of protein and protein-protein interaction data within public biological databases is described. For a comprehensive and consistent way of searching and analysing integrated protein and protein-protein interaction data, the InSilico Proteomics (ISP) project has been initiated. Its three main objectives are (1) to provide an integrated knowledge pool for data investigation and global network analysis functions for a better understanding of a cell's interactome, (2) employment of public data for plausibility analysis and validation of in-house experimental data and (3) testing the applicability of Microsoft's .NET architecture for bioinformatics applications. Data integrated into the ISP database can be queried through the Web portal PRIMOS (PRotein Interaction and MOlecule Search) which is freely available at http://biomis.fh-hagenberg.at/isp/primos.

1 Introduction

Public biological databases are the major resources for researchers in the field of molecular biology. A well-known problem is the distribution of information among multiple databases. The problem rests upon the focus on different types of data, as for example, sequence and structural information, function or protein domain descriptions, molecular interaction or publication data. In order to get a comprehensive view of certain research questions, the researcher has to consult multiple biological databases.

In cooperation with researchers of the dermatologic field a system was developed to handle this problem in the research field of protein-protein interaction analysis. Therefore, publicly available protein annotation data and information about protein-protein interactions were integrated into a data warehouse.

The InSilico Proteomics (ISP) system is accomplished within the infrastructure setup project BIOMIS (BIOMedical Information Systems). Its purpose is to build software systems with Microsoft .NET technology in collaboration with a partner of the life science research field.

The .NET Framework has been chosen because of the following reasons: (1) availability of a huge foundation class library, (2) native XML support, (3) support of many programming languages that can interchange and directly use developed libraries, (4) availability on multiple operating system platforms, (5) concepts for an easy development of both Desktop and Web applications.

The main task of the ISP system is to build an application for supporting molecular biologists in data investigation. In the proteomics field there exist a lot of databases that cannot be analyzed at once. Hence, the ISP system should integrate the most important of these databases and provide a common search interface. Additionally, this huge data pool should be reused for plausibility analysis and validation of experimental data.

Importing data from multiple sources is a very challenging task. Each database provides its own levels of detail and complexity, data formats, and methods to access the data. Accessing and transforming data from different sources takes 70-80 per cent of the time of creating an integrated biological data warehouse (see [19]). One might ask why — despite this expected effort — we started this job although these data are already accessible to researchers? In addition, solutions to access multiple data sources at a time, like the Sequence Retrieval System (SRS) [5], already exist. There are various reasons pointing in favour of local data warehousing according to [24]. With a local copy of public biological data one can

1. perform complex analyses and queries.
2. increase efficiency since no remote access to potentially slow or non-responding external sources is necessary.
3. integrate multiple databases into one database and cumulate pieces of information.
4. clean data more carefully than during on-the-fly data access.
5. perform calculations with intense data access, which cannot be done online due to fair-use and time limitation when accessing remote resources.

2 Methods

Caused by the increasing number of biological experiments producing vast amounts of experimental data, biological databases became an ever-growing field within the last decades. In [7], Galperin mentions a growth of 139 molecular biology databases within one year leading to 858 public databases stated in this report.

A major problem of biological information representation is caused by the fact that it is hard to describe and not static. Especially the different aspects of information representation can be a pitfall for efficient data processing and data integration. Because of the diverse development advanced by different institutions, an intricate combination of data formats and ways of accessing biological data evolved.

2.1 Data Formats

Automated analysis and integration of data stored in public biological databases has become more and more interesting during the last decade. This is mainly due to decreasing storage and computing costs, but also because of the increasing data amounts created in countless experiments all over the world and stored in public biological databases. Text files (often called flat files) are very convenient for human readers because the data is formatted in a way they are familiar with. For an efficient automated analysis, however, the need for data formats that structure data in a way to be easily processed by computers emerged. ASN.1 (Abstract Syntax Notation One) enables the representation of both, data and structure within a single file. NCBI uses ASN.1 for the storage and retrieval of nucleotide and protein sequences, structures, genomes, and MEDLINE records [16].

Within the last years, more and more data are structured using the eXtensible Markup Language (XML) that allows the specification of a data schema. The file content can be automatically checked by schema-checking XML parsers. Automatic parser generation based on a given schema and XML querying techniques like, e.g., XQuery [22] allow a powerful handling of even large datasets. In the field of bioinformatics, the pros and cons of XML usage and applications of developed data schemas have already been discussed in the literature (cf. [1], [10] and [6], [9]).

Another application of XML-based data formats is settled in the field of biological data exchange standards. Especially in the proteomics field, the HUPO Proteomics Standards Initiative (PSI) combines strong community efforts for proteomics data representation to facilitate comparison, exchange, and verification (see [13]). Divided into several working groups, recommendations for protein-protein interaction, mass spectrometry, general proteomics, and proteomics informatics data are developed. Within the ISP system, especially the PSI Molecular Interaction (MI) standard described in [12] is of high interest because it reduces implementation efforts by means of a consistent format for molecular interaction data. This reduction is achieved by commitments of public interaction databases to release their data following the PSI-MI standard. Additional efforts for sharing curation tasks and record exchange are bundled within the International Molecular Exchange Consortium (http://imex.sf.net).

2.2 Types of Data Access

The most commonly used approach to access biological data are Web interfaces providing a convenient way for querying and browsing public biological data. Because these databases have been created by different universities or companies, every database provides its own style of data representation. Hence, this diversity of user interfaces led to preferences of molecular biologists for the usage of a specific database. In general, the main focus of Web interfaces is to serve a human user's requirements; i.e., their usage and appearance is optimized in this way. For automated access, Web interfaces cannot be used efficiently.

In contrast to Web interfaces whose main mission is to provide search and browsing functionality, sometimes users want to download and process full data sets to do further analysis. One of the most common ways of data exchange in the internet is the *File Transfer Protocol (FTP)*. Data is provided on an FTP server and FTP client applications running on the local machine can download the required information.

When it comes to the need of fetching a specific data entry from a public database, it would be too much effort to download the whole set of data files, scan these files for the desired entry, and apply the required action on this entry. Thus, public databases provide other types of access methods that allow retrieval of specific entries or queries within the databases. NCBI's eUtils [18], e.g., provide access to NCBI databases using a standardized URL syntax. Query strings, options, and the output format are set through URL parameters, which makes this approach very convenient for automated access. In order to provide reliable systems, service providers like NCBI impose strict access limitations (e.g. no more than one request every three seconds should be made).

Within the last years, the emerging technology of Web services got more and more popular. In contrast to proprietary remoting methods like .NET remoting or remote procedure calls (RPC), Web services use open, XML-based standards and protocols to exchange data between communicating systems. As Web services are supported by modern programming languages, there are almost no limitations of implementing client programs. In the field of bioinformatics, more and more biological databases provide Web service interfaces for accessing their data. NCBI provides Web services similar in syntax to eUtils for data access. UniProt Web services [17] can be used for a number of tasks, like data retrieval or sequence analysis. The benefit of using Web services is that developers do not have to care about data transmission and data parsing since the data is received in a way that can directly be processed by applications.

Another level of access beyond the above mentioned types is provided by public data integration portals and databases. Since several years, efforts have been put into the implementation of such services, like SRS or Biozon [3]. The most widely used database integration system, SRS, provides functionality to search across public, private or even licensed databases. Data connectivity is achieved through the definition of a script describing the structure of the data. As these systems do not necessarily create or discover inter-database connections, they are often described as user interface integration rather than data integration services.

3 Results

We created a system that integrates multiple databases in the field of proteomics and reuses the data within complex analysis methods. In the following we describe the four main architectural layers of the ISP system and their functionality.

3.1 Data Import Layer

The implementation of the integration approach is realized within the data import layer of the ISP system. Information domains relevant for the ISP system are (1) protein annotation databases like Swiss-Prot, TrEMBL [4], and the NCBI Protein database [23], (2) protein-protein interaction databases such as BIND [2], IntAct [11], MINT [26], and DIP [25] and (3) additional annotation information like NCBI Taxonomy database [23], GeneOntology vocabulary [8] and PSI-MI vocabulary [12].

Data Integration Basically, there had been two major integration challenges. First, information belonging to a certain domain and originating from different databases, like e.g. protein

data, had to be integrated into a common schema. Second, data belonging to complementary information domains like proteins and protein-protein interactions also had to be combined. These two approaches — vertical integration for the aggregation of semantically similar data — and horizontal integration for the composition of semantically complementary data have already been discussed in detail in [21].

The most efficient method to access those heterogeneous sources listed was to use XML-formatted data wherever possible. Within the Microsoft .NET technology XML parsers can be easily created with little effort for a given schema. The main advantage of this approach is that one does not need to care about correctly interpreting the data, since they are automatically parsed. As long as the files are structured correctly according to the schema they reference, this is the easiest way of reading files for further processing. Any change in the XML data schema needs to be reflected in the parser, so revision notifications from the public institutions hosting remote sources are essential for keeping the parser framework up-to-date. For flat files it is more time-consuming to create parsers, since the structure of the data is not specified in the data file itself. Often the meaning of each data block has to be looked up in separate documentation files which sometimes are out-dated or in the worst case do not exist at all.

Table 1 gives an overview of the currently imported data in the ISP database. For every public data source that has been integrated, the number of imported entries, the utilized data formats, and access types within the ISP Data Import Layer are mentioned. As already stated, XML formatted data has been preferably used. Especially in the field of protein-protein interaction data, the availability of the common data standard (PSI-MI) alleviated the import process. During initial imports, data have been automatically downloaded using FTP, parsed, transformed wherever necessary, and afterwards stored into the ISP database. Data updates also mainly rely on FTP access, but for updates of specific data entries URL-based methods or Web service routines have been used.

Data Source	Nr. of Entries	Data Format	Access Type
Swiss-Prot	227,080	XML	FTP, Web services
TrEMBL	2,570,041	XML	FTP, Web services
NCBI Protein	234,912	XML	FTP, eUtils
BIND	72,656	flat file	FTP
IntAct	70,675	PSI-MI	FTP
MINT	68,478	PSI-MI	FTP
DIP	planned	PSI-MI	FTP
PSI-MI controlled vocabulary	676	XML	FTP
GeneOntology	21,096	XML	FTP
NCBI Taxonomy	306,340	XML	FTP, eUtils
PubMed	317,271	XML	FTP, eUtils

Table 1: Overview of the data sources and the combinations of format and access methods used by the Data Import Layer of the ISP system (Database statistics as of July 2006).

Data Updates Once data is integrated into the data warehouse the next challenge approaches, since the local data pool has to be kept up-to-date with its sources. Not only *how* but also *when* updates are to be performed is a big problem to be solved, because updating the local data pool takes a certain amount of time and hardware resources. In order to minimize administrative

work for updates, the ISP database changed to a quarterly release cycle. Releases are prepared within a semi-automatic process initiated by a database administrator.

Another aspect that would ease data updates are delta files (files containing changes since the last revision). Unfortunately, these files are sparsely provided, but the whole data is available in one single file. To handle such files in an adequate period of time the new version of the data file is pre-processed to filter out those pieces of information that need to be updated in the local data warehouse.

3.2 Data Storage Layer

The ISP data storage layer is responsible for data management and can be divided into two major parts connected by an interlinking area (see Figure 1). The public data pool holds data integrated from publicly accessible sources like protein databases, protein interaction resources, publication information, and controlled vocabulary to further describe a basic entry. The private part of the database is a basic management system, which stores experimentally found protein-protein interaction data from a lab. Its main focus lies in administrating and validating protein interactions observed in laboratory experiments. Laboratory results are associated with public information, e.g. homologs (i.e. orthologs and paralogs) or fragments of a protein are identified. In both schema parts analysis and calculation results are stored.

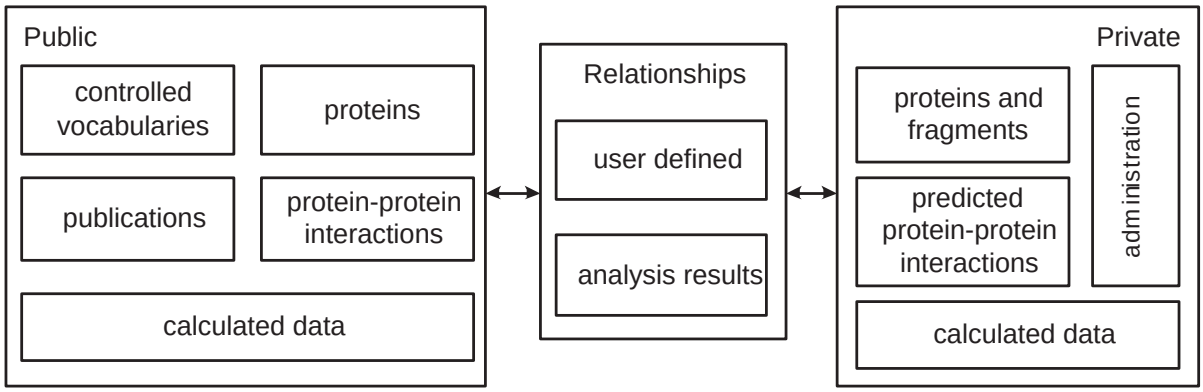

Figure 1: Core database areas of ISP system.

The public data pool's schema is based on the open source database schema BioSQL [14]. The main focus of BioSQL lies in the storage of biological sequence data and annotations. Although being well suited for storing protein information, the schema had to be extended to handle protein-protein interaction data. This part of the database is mainly based on the database schemas proposed by IntAct and MINT (see [11] and [26]). Even though these schemas provide the possibility of storing information about interaction partners, they have been merged with the BioSQL schema to combine the strengths of both approaches. Therefore, it has been necessary to identify shared or overlapping entities and merge these two domain schemas to a global ISP database schema.

To reconcile the variety of data it is advisable to assign consistent descriptions to proteins and protein-protein interactions through controlled vocabularies and ontologies. Fortunately, there are already a couple of vocabularies defined in the bioinformatics area so it was not necessary to define a proprietary vocabulary within the ISP system. One approach to represent

a classification is NCBI's Taxonomy database, where among other things, a unique number is assigned to each organism. Identifying an organism by this number avoids the problem of misspelling or the usage of synonymous names of one and the same organism. Gene and gene product attributes are classified through a vocabulary of terms defined by the GeneOntology consortium. In the field of protein-protein interactions, the vocabulary defined by the PSI-MI group paves the way to a common nomenclature.

3.3 Data Analysis Layer

The ISP database not only contains integrated public information but is extended by analysis and calculation results. One of these analysis results is the grouping of redundant protein or protein-protein interaction entries in the database. The problem that arose during the integration of protein and interaction data was the aspect of missing entry cross-links between similar protein or interaction entries coming from different or even the same public biological databases. Hence, protein entries have been grouped according to their identical amino acid sequence and organism information. Within this definition 8 per cent of over 3 million proteins in the ISP database were combined in groups. The biggest protein group formed by this approach tied together 682 protein entries. Furthermore, protein-protein interaction data have been grouped in respect of equal interaction partners, equivalent experimental methods and publications. As these parameters do not reflect different experimental parameters (like a dissociation constant) which are not available for all interactions the grouping leads to similar, not mandatory equal, interaction observations. According to these requirements 21 per cent of the current 200.000 interactions in the ISP database where identified to be similar, where the biggest group of similar interactions bundles 59 interactions.

As an example, the *tumor suppressor protein p53* is represented by 4 protein entries in the ISP database, three of which were imported from NCBI (NP_000537, XP_008679, CAA42627), one coming from TrEMBL (Q2XN98). Having all pieces of information from multiple protein entries describing the same protein at a glance is a big advantage, since they complement one another. The cumulated information of every protein group has not been physically merged within the ISP database, but can be displayed within the Web interface (see Figure 2). In addition, detailed information of every group member can be requested.

3.4 Data Presentation Layer

The ISP database has two visual user interfaces: The search platform PRIMOS (PRotein Interaction and MOlecule Search) accesses the integrated public data, whereas the laboratory application PRIMOSLab is a management system that links experimental data to public information and provides analysis methods to lab staff.

The PRIMOS engine (http://biomis.fh-hagenberg.at/isp/primos/) is a Web-based search platform to query public data integrated into the ISP database. It offers basic search options like name or public identifier search. However, its main focus lies in providing more complex search approaches that give an overall picture of an area of interest. Three of which are implemented to date are: (1) Identifying hub proteins of network complexes, which is of interest for interaction detection. Those highly interacting proteins represent cores of molecular networks and are the most useful baits [15]. They are of major interest for biologists, because

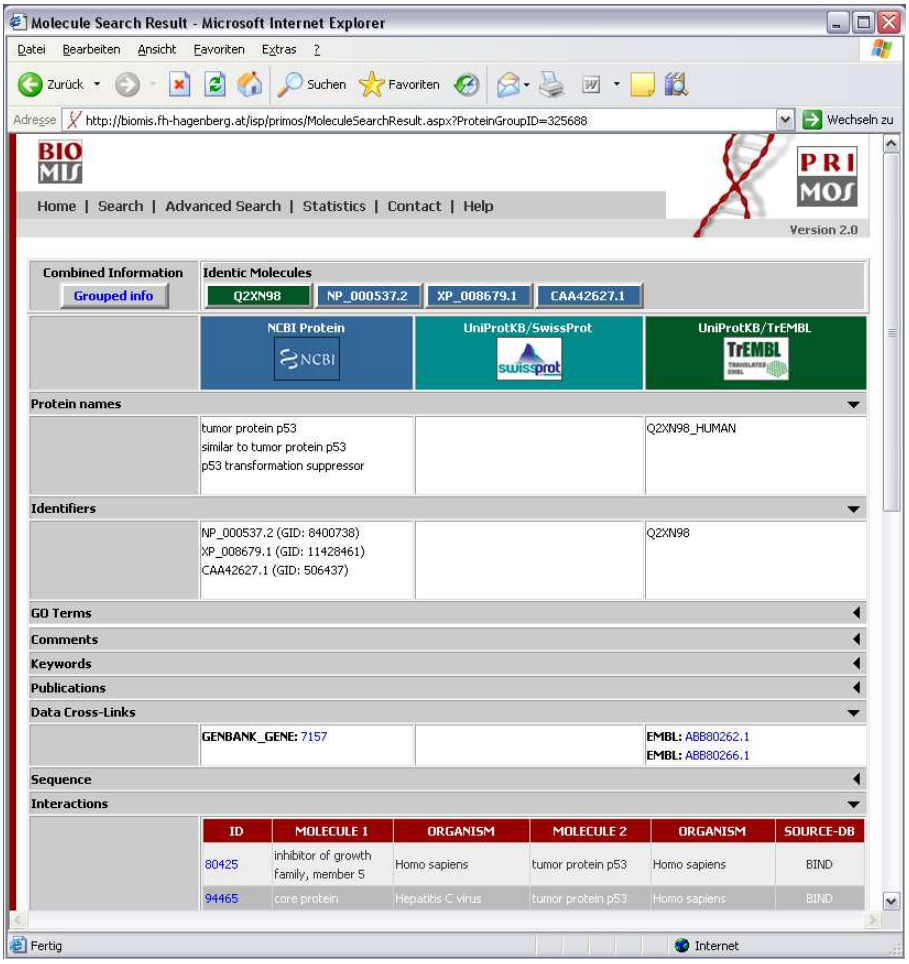

Figure 2: Cumulated display of tumor protein p53.

mutation of such proteins has a great impact on metabolic pathways due to their widespread connectivity. (2) Interactions between proteins of different organisms may be an indication for host-pathogen reactions. Finding such organism cross-talks is also supported in PRIMOS. (3) Well-researched, confirmed interactions can be easily found by searching for interactions observed within different experimental methods. This is important since experimental methods like yeast two-hybrid are known to produce certain false positives [20]. Additionally, it is acknowledged that many biological sources contain a large number of errors [24]. In the following these advanced search functions are described in more detail.

Network Hub Search Network hubs can be identified by the number of interactions the hub protein is involved with. If a certain organism is of interest the hub search can be narrowed to this organism. As an example, the search for all proteins in Homo sapiens which have at least 50 interactions with other proteins was performed using PRIMOS. The results show the existence of 71 currently known small networks in Homo sapiens, with well known representatives such as the already mentioned *tumor suppressor protein p53*, *Nuclear factor NF-kappa-B p100 subunit* or the *transforming growth factor, beta receptor I precursor*. The information is returned in list-format; furthermore, an interactive network visualization is available in a protein's detail view.

Organism Cross-Talk Identification Identifying the cross-talks between organisms is of particular importance for the investigation of pathogen and host interactions. It supports the understanding of pathogen replication and pathogenesis processes. This provides an essential foundation for the development of safe and effective therapeutic and preventive strategies to combat pathogens. PRIMOS yielded 1,274 interactions between proteins of Homo sapiens and Human immunodeficiency virus 1. A search for cross-talk interactions between proteins of Homo sapiens and Hepatitis C virus revealed 72 interactions. Additionally, this search option can be narrowed by the selection of a distinct experimental method the interaction was found with. Since many different techniques exist to identify protein-protein interactions, the user could be interested in the results of one class of experimental analysis providing certain pieces of information of the interaction. This approach can be helpful for the comparison of results generated by different experimental techniques. Not only organism-crossing interactions can be of interest: By selecting the same organism twice, all interactions within this organism observed by an experimental method can be revealed.

Confirmed Interaction Search It is well known from large scale protein-protein interaction analysis, that the data produced are often a blueprint of the investigated system, with a high percentage of false positives. A significant part of these false positives appear due to the experimental method itself. For example, the yeast two-hybrid system is known to produce certain classes of such false positives [20] by placing quality behind quantity of data. System immanent false positives always exist and have to be eliminated by additional experimental procedures. Therefore, a search routine for the identification of trustable results, meaning results confirmed by a number of methods is included in the Web interface. The investigator determines the quality of the interaction data either by setting the number of confirmations for protein-protein interactions or by defining the experimental methods the interactions have been identified with. E.g. all protein interactions in Human immunodeficiency virus 1 (HIV-1) were searched for interactions confirmed by at least four independent experiments. PRIMOS found 413 protein-protein interactions such as our well-known human *tumor supressor protein p53,* which is interacting with the *Nef gene product* of HIV-1.

PRIMOSLab is the restricted laboratory extension of the public search engine PRIMOS. This part of the system in the current state is a prototype application for administration and in-silico validation of experimental data. It stores experimental results like protein fragments identified and interactions observed. The user input, which contains sensitive data, is stored in the private part of the ISP database not accessible to PRIMOS users. The in-silico validation methods are based on comparative proteomics and use public information like functional annotation and sequence data as a knowledge base for the validation process. To date no structural protein information is used. A demo version of the application prototype is available at http://biomis.fh-hagenberg.at/isp/primoslabdemo/.

4 Conclusion

In this paper we described the basic architecture of the InSilico Proteomics System. The ISP system integrates data from standard databases of the proteomics research area. The database

is kept up to date by a semi-automatic updater framework. During the build-up phase of the ISP database we were faced with the challenges of data integration. We realized that combining data from different sources which are available through a multitude of data formats, data access methods and in different levels of detail requires a good strategy and always constitutes a compromise.

The decision to store the data locally was made because of the main requirement of the ISP system to perform intense data access. In order to keep their systems available for the audience public resources have to enact access limitations, which did not allow us to access data in a frequency necessary for some ISP analysis methods. Out of this decision we take the advantage not to depend on external data sources directly.

Goal of the ISP system is to serve as knowledge base for validation and plausibility analysis of in-house data. A separate system application is responsible for managing and administrating laboratory-internal data and uses the public data pool for comparative validation methods. The search engine PRIMOS is a publicly available Web-interface for querying the ISP database. The advanced search options are an important feature to get an insight into functional networks built by protein-protein interactions. Investigations in the field of drug discovery might rise the question whether there are any proteins in an organism interacting with a certain number of other proteins, since such core proteins have an enormous impact on the functionality of molecular pathways. Additionally, analyzing interactions between different organisms provides important knowledge about host-pathogen reactions.

Additionally to public data, the ISP database contains analysis and calculation results, where the grouping of protein and protein-protein interactions, respectively was mentioned as an example. This way the researcher is supported in data investigation because pieces of information spread in multiple data entries can be cumulated. Furthermore, the true number of known interactions a protein participates in can be discovered when eliminating redundant entries. Hence, the grouping approach is a technique to reduce redundancy and complexity in the ISP database.

We built our system with the support of Microsoft .NET technology. Even though there are not yet resources available in .NET for processing bioinformatical requests, this framework is appropriate to accomplish the following tasks: Parsing XML data efficiently is fully supported so one can focus on further data processing. The same level of support is given for Web services, which are one of the most popular methods for accessing remote data. This is the case because the data can be interpreted beyond different technologies. Presenting the data in a Web application can be realized within the same technology since the .NET framework supports the development of interactive Web pages, where a Data Access Layer can be directly linked to the Data Presentation Layer.

5 Acknowledgements

The InSilico Proteomics System is carried out at the Research Center of the Upper Austrian University of Applied Sciences (UAS) in Hagenberg (AUSTRIA) as an FHplus infrastructure setup activity. FHplus is a research program, funded by two Austrian ministries and organized by the Austrian Research Promotion Agency (FFG) to setup and enhance R&D capacity and competence in Austrian UAS.

References

[1] Archard F., Vaysseix G. and Barillot E. XML, bioinformatics and data integration, *Bioinformatics*, 17, 115-125, 2001.

[2] Bader G.D., Betel D. and Hogue C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31, 248-250, 2003.

[3] Birkland A. and Yona G. BIOZON: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics*, 7, 70, 2006.

[4] Boeckmann B., Bairoch A., Apweiler R., Blatter M.C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S. and Schneider M. The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res*, 31, 265-370, 2003.

[5] Etzold T., Ulyanov A. and Argos P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*, 266, 114-128, 1996.

[6] Feny D. The Biopolymer Markup Language, *Bioinformatics*, 15, 339-340, 1999.

[7] Galperin, M.Y. The Molecular Biology Database Collection: 2006 update, *Nucleic Acids Res*, 34, D3-5, 2006

[8] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Res*, 32, D258-261, 2004.

[9] Gilmour, R. Taxonomic markup language: applying XML to systematic data, *Bioinformatics*, 16, 406-407, 2000.

[10] Guerrinie VH. and Jackson D. Bioinformatics and extended markup language (XML), *Online Journal of Bioinformatics*, 1, 1-13, 2000.

[11] Hermjakob H., *et al*. IntAct - an open source molecular interaction database, *Nucleic Acids Res*, 32, D452-D455, 2004.

[12] Hermjakob H., *et al*. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data, *Nat Biotechnol*, 22, 177-83, 2004.

[13] Kaiser J. Public-Private Group Maps Out Initiatives, *Science*, 296, 827, 2002.

[14] Lapp H. *et al*. Open Biological Database Access, *Available Online*, http://obda.open-bio.org, 2005.

[15] Lappe M. and Holm L. Unraveling protein interaction networks with near-optimal efficiency, *Nat Biotechnol*, **22(1)**, 98-103, 2004.

[16] National Center for Biotechnology Information. ASN.1 Summary, *Available online*, http://www.ncbi.nlm.nih.gov/Sitemap/Summary/asn1.html,.

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

[18] Sayers E., Wheeler D. Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils), *Available Online*, http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/coursework/chapter_eutils.pdf, 2006.

[19] Schonbach C., Kowalski-Saunders P. and Brusic V. Data warehousing in molecular biology, *Brief Bioinform*, 1, 190-198, 2000.

[20] Serebriiskii IG. and Golemis EA. Two-hybrid system and false positives. Approaches to detection and elimination, *Methods Mol Biol*, 177, 123-124, 2001.

[21] Sujansky W. Heterogeneous Database Integration in Biomedicine, *Journal of Biomedical Informatics*, **34**, 285-298, 2001.

[22] W3C XQuery 1.0: An XML Query Language, *Available online*, http://www.w3.org/TR/xquery/, 2005.

[23] Wheeler, D.L., *et al*. Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res*, 33, D39-45, 2005.

[24] Wong L. Technologies for integrating biological data, *Brief Bioinform*, **3**, 389-404, 2002.

[25] Xenarios I., Rice D.W., Salwinski L., Baron M.K., Marcotte E.M. and Eisenberg D. DIP: the Database of Interacting Proteins, *Nucl. Acids. Res*, 28, 289-291, 2000.

[26] Zanzoni A., Montecchi-Palazzi L., Quondam M., Ausiello G., Helmer-Citterich M. and Cesareni G. MINT: a Molecular INTeraction database, *FEBS Letters*, 513(1), 135-140, 2002.