Ongoing development of Two-Dimensional Polyacrylamide Gel Electrophoresis Data Standards

Veerasamy Ravichandran[1], Joshua Lubell[2], Gregory B. Vasquez[1], Peter Lemkin[3], Ram D. Sriram[2], Gary L. Gilliland[1]

[1]Biotechnology Division, [2]Manufacturing Systems Integration Division, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899

[3]Laboratory of Computational and Experimental Biology, National Cancer Institute at Frederick, Frederick, MD 21702

Running Title: Unifying 2-D PAGE data

Correspondence: Dr. V. Ravichandran, Center for Advanced Research in Biotechnology of the University of Maryland Biotechnology Institute and the National Institute of Standards and Technology, 9600 Gudelsky Drive, Rockville, MD 20850, USA

Email: vravi@nist.gov

Fax: +1-301-738-6255

Abbreviations: 2-D PAGE, two-dimensional polyacrylamide gel electrophoresis; DTD, document type definition; HTML, hyper text markup language; ISO, International organization for standardization; SVG, scalable vector graphics; TWODML, two-

dimensional electrophoresis markup language; XML, extensible markup language; XSL, Extensible style sheet language

Keywords: 2-D gel electrophoresis, Data standards, Interoperability, Markup language, Proteomics

Summary

Two-Dimensional Polyacrylamide Gel Electrophoresis (2-D PAGE) is a common method used to fractionate and identify proteins. Many technical advancements including automation towards this method lead to accumulation of data. As 2-D PAGE plays a major role in proteomics research, it is necessary to compare data from numerous gels, and many of theses samples are unique and cannot be reproduced. Currently, no data standards exist for 2-D PAGE data to systematically establish the correlation between data from different experiments to allow meaningful comparisons. This raises the difficulty in data exchange and interoperability between different data resources even though they may have some data in common. Thus, data interoperability requires the development of standard data definitions and transfer protocols. Such standards for databases and data reporting can be applied to 2-D PAGE technology, a critical method employed in proteomics efforts. Comprehensive, structured information about the 2-D PAGE data will aid the deeper understanding of a particular protein. The adoption of common standards and ontologies for the management and sharing of 2-D data is essential and will provide immediate benefit to the proteomics community. Hence, we perceive a role for the National Institute of Standards and Technology (NIST) and other similar standards and measurement organizations in facilitating this development. In this paper we present an approach toward standardizing 2-D PAGE data in support of developing a globally relevant proteomics consensus in order to provide more efficient database querying and data comparisons through the establishment of the necessary definitions and interdisciplinary reference fields for both the 2-D PAGE community, particularly in the proteomics area, and the clinical and experimental biological research

communities, in general. This article covers the need for unifying the 2-D PAGE data through a common data repository, and its usefulness in data standards and data interoperability.

Introduction

Two-Dimensional Polyacrylamide Gel Electrophoresis (2-D PAGE) has been an important tool for biological research for decades, but with the advent of proteomic research there has been a surge in both the number of users and the number of gels that are being run. The biological community marks the completion of the Human Genome Project's major goal in 2003: complete, high-accuracy sequencing of the human genome, which led to the discovery of genes coding for the production of tens of thousands of proteins [1, 2]. The current focus is on the gene products, specifically proteins, and the overall biological systems in which they act, creating the emerging fields of systems biology and proteomics. The free, widespread availability of a wide variety of data beyond human genome sequences - sequence variation data, model organism sequence data, organelle specific data, expression data and proteomic data, to name a few - is starting to provide the means for scientists in all disciplines to better-design and interpret their laboratory and clinical experiments, hopefully accelerating the pace of biological discovery. Proteomics research is resulting in enormous amounts of data, orders of magnitude larger than the data generated in genomics studies, making the effective and efficient management of data essential. Having such a rich source of information is proving invaluable to scientists, whose findings should, in time, lead to improved strategies for the diagnosis, treatment and prevention of genetic diseases [3, 4].

However, it has become increasingly clear that simply generating the data is not enough; one must be able to extract from it meaningful information about the system being studied. Biological system data holds phenomenal promise for identifying the

mechanistic basis of organismal development, metabolic processes, and disease, and it can confidently be predicted that bioinformatics research will have an additional impact on improving our understanding of such diverse areas as the regulation of gene expression, protein structure determination, comparative evolution, and drug discovery [5-9]. Of central importance to the optimal utilization of proteomics data, particularly 2-D data is the development of a unified infrastructure to facilitate collecting, storing, retrieving and querying data, regardless of the technology used to generate it.

Need for Data Standards

Data standards are essential because they permit cooperative interchanges and querying between diverse, yet dissociated databases. The ability to interchange data in a seamless manner becomes critically important, and the economic benefits of data interoperability standards are immediate and obvious.

Evaluated/Annotated data for Data exchange

Lack of data standards is the Achilles heel of data interoperability. The lack of integration, implementation and use of standards are barriers to the delivery of optimal biological data. Even with the dramatic increase in the volume of proteomics data, innovation will be constrained unless technical advances are made in producing critically evaluated data and integrating data sources through data management, mining, and integration techniques. In the context of time-to-completion pressures and volumes of data, the research community needs certified models that can derive “best” recommended values from critically evaluated experimental data and validated and benchmarked predictive methods for any real or proposed measurement. These virtual measurement systems could generate data suitable for immediate use in commercial, scientific, and

regulatory applications. Also needed are effective data management standards and techniques (e.g., quality, traceability, or uncertainty estimates) for gathering, integrating, and maintaining information about data accessed from diverse sources [10]. Comparison of 2-D PAGE data between experiments and laboratories is essential, particularly as this data is utilized in clinical diagnostic settings. Accepted data standards for 2-D PAGE experiments would be a valuable asset for the study of diseases, especially when samples are rare and often unique. With respect to querying across biological databases, required information for a particular protein or family of proteins can be difficult to obtain from heterogeneous resources because of lack of data standards, leading to data compatibility problems. Thus, accurate and comprehensive automated information exchange between these resources is very limited.

Although public proteomics data resources are highly informative individually, the collection of available content would have more utility if provided in a standard and centralized context and indexed in a robust manner. Use of controlled vocabulary is already facilitating analysis of high-throughput data derived from DNA microarray experiments. The potential impact of improved interoperability derives from the fact that information and knowledge management systems have become fundamental tools in a broad range of commercial sectors and scientific fields [11-13]. The adoption of common standards and ontologies for the management and sharing of proteomics data is essential [14]. Even though, there exist many biological data exchange formats (Table 1) [14], a well documented and annotated data with easy exchangeable data format, such as an

extensible markup language (XML) format, would help in data mining, annotation, storage, and distribution.

XML for biological data exchange

Extensible Markup Language (XML) was defined by the XML Working Group of the World Wide Web Consortium (W3C: http://www.w3.org/XML). XML is a markup language for documents containing structured information. Structured information contains both content (words, pictures, etc.) and some indication of what role that content plays (for example, content in a section heading has a different meaning from content in a footnote, which means something different than content in a figure caption or content in a database table). Almost all documents have some structure. A markup language is a mechanism to identify structures in a document. The XML specification defines a standard way to add markup to documents. XML is a simple, very flexible text format, playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere [15]. Because an XML document so effectively structures and labels the information it contains, the web browser can find, extract, sort, filter, arrange, and manipulate that information in highly flexible ways. XML has been designed for ease of implementation and for interoperability with the World Wide Web.

The XML definitions consists of only a bare-bones syntax. When an XML document is created, rather than use a limited set of predefined elements, the data elements are created and subjectively assigned names as desired – hence the term extensible in Extensible Markup Language. Therfore, XML can be used to describe virtually any type of document, and fits ideally with the requirements for the complex and diverse biological

data integration. XML thus provides an ideal solution for handling the rapidly expanding quantity and complexity of information that needed to be put on the web. Hence, a common language, like XML, should therefore offer power, scalability, adoptability, interoperability, flexibility with different data types. In order to enhance the interoperability between diverse data, adoption of a universal data exchange language, like XML, to exchange the annotated data would be useful. The number of applications currently being developed by biological communities that are based on, or make use of, XML documents is truly amazing. This is to facilitate the writing and exchange of scientific information by the adoption of a common language in XML (Table 2).

Markup Language for Electrophoresis Data

Although there has been rapid progress in establishing standards for genomic sequence data as well as DNA microarray data, it is unlikely to be the case with proteomic data. Human Proteome Organization's (HUPO) Protein Standards Initiative (PSI) aims to facilitate standards for data representation in proteomics. Currently, the PSI is focusing on developing standards for two key areas of proteomics; mass spectrometry and protein-protein interaction data that will be XML based. We propose here a common language for sharing electrophoresis experimental data, Two-dimensional Electrophoresis Markup Language (TWODML) that is based on the XML. Given the amount, complexity of data associated with a single set of electrophoresis experiment, we believe that XML is the most suitable method to describe electrophoresis data. The goal of the TWODML is to:

1) Gather, annotate, and provide enough information that may be reported about an electrophoresis based experiment in order to ensure the interoperability of the results and their reproducibility by others;

2) Help establishing public repositories and data exchange format for electrophoresis based experimental data;
3) Eliminate barriers to data exchange between the electrophoresis data, and permit the integration of data from heterogeneous sources; and
4) Leverage low-cost XML-based technologies such as XSLT (Extensible Style sheet Language Transformation) and SVG (Scalable Vector Graphics).

The first step in data interoperability is to enforce standards for the electrophoresis data. A much bigger challenge is arriving at what descriptors constitute a 'required/minimum acceptable' set with respect to different types of parameters. The minimum information necessary from any 2-D PAGE experiment is that associated with the experimental details, in order to ensure firstly the reproducibility of the experiment, and secondly the interoperability of the results. An electrophoresis data standards committee can define a general-purpose set of TWODML elements, together with a document structure, to be used for a particular class of documents. By defining the vocabularies in a standard format (e.g., the experimental sample source) the resulting uniformity may permit comparison of data between different systems (microarray data, macromolecular data, etc.). Hence, in the case of 2-D PAGE, the common data elements and data definitions for the required information were outlined.

The following data elements should be collected in association with their required data categories: sample source, detail about the protein, experimental detail, sample preparation, sample loading, sample separation condition, sample separation,

experimental analysis, data analysis and author information (Fig. 1) [#]. For example, the sample source information should contain the following tentative data items: the source record, which specifies the biological and/or chemical source of each molecule in the entry. Sources should be described by both their common and scientific names. Two types of sources will be grouped: the natural source (Fig. 2) and genetically modified (recombinant) source (Fig. 3). A part of recombinant source TWODML model data conforming to the schema fragment shown in figure 4 might appear as in Figure 4. For example, the gene used in the recombinant sample source is described as below:

```
<TWODML>
  <SAMPLE_SOURCE>
   <RECOMBINANT_SOURCE>
      <GENETIC_MATERIALS>
        <GENE>
              <NAME>SNAP-23</NAME>
              <PROTEIN_NAME>Snaptosome-associated Protein of 23 kDa
              </PROTEIN_NAME>
              <SOURCE>
                 <ORGANISM_SCIENTIFIC>Homo sapiens
                 </ORGANISM_SCIENTIFIC>
                 <ORGANISM_COMMON>Human
                 </ORGANISM_COMMON>
                 <CELL_LINE>Raji - human B lymphocyte (Burkitt's Lymphoma).
                              ATCC number: CCL-86
                 </CELL_LINE>
              </SOURCE>
              <GENETIC_VARIANCE>Amino acid 23 is changed from Ser to Ala
                                  (Ser23Ala)
              </GENETIC_VARIANCE>
              <ORIGIN>
                 <NAME>Dr. Roche</NAME>
                 <ADDRESS>National Cancer Institute, NIH</ADDRESS>
                 <CONTACT_INFO>pr17m@nih.gov</CONTACT_INFO>
              </ORIGIN>
```

[#] Figures 1 through 4 were created using XMLSPY's schema designer. Rectangular boxes in these diagrams represent XML elements. The octagonal symbols are sequence compositors. Thus Figure 1 indicates that a TWODML root element shall contain a SAMPLE_SOURCE element, followed by an EXP_DESIGN element, and so on.

```
        <MORE_INFO>GeneBank/EMBL Data Bank accession number:
                   U55936
        </MORE_INFO>
</GENE>
```

In this partial TWODML document, data elements are marked by enclosing sets of angle brackets. Standardized values for well defined data elements are embedded within the elements. In addition to well defined syntax described above, each TWODML document also carries information about: data relationships, data types, range restrictions, interdependencies, exclusivity, units, and methods. Free, open-source software is available to validate and parse the data files.

As in the natural source, data items in the genetically-modified category record details of the source from which the sample was obtained. Associated data for this category include: the gene modified in the source material for the experiment, the genetic variation (transgenic, knockout), the system used to express the recombinant protein, and the specific cell line used as the expression system (name, vendor, genotype, and phenotype). Data items in the natural source category will record details of the sample source. Associated data for this category will include: the common name of the organism and its scientific name, the source condition (normal, disease), any genetic variation, sex, age, organ, tissue, cell, organelle, secretion, and cell line information. The cell line and strain should be given for immortalized cells when they help to uniquely identify the biological entity studied.

The TWODML application is being defined by creating an XML schema which defines and names the elements and attributes that can be used in the document, the order in

which the elements can appear, and constrains the values of elements and attributes. A variety of schema languages are available for XML. The oldest and simplest of these schema languages is the Document Type Definition (DTD) [16]. Although DTDs enjoy strong software support, they are inadequate for representing strongly typed or context-sensitive information. Two newer XML schema languages, the W3C's XML Schema Definition Language [17] and ISO's RELAX NG [18], address these shortcomings of DTDs. Therefore, we plan to specify the TWODML schema using one or both of these languages. Our application will also include one or more Extensible Style sheet Language (XSL) style sheets. These style sheets will enable electrophoresis XML data to be transformed into other useful formats such as HTML (Hyper Text Markup Language) and SVG.

Data Repository or Electrophoresis Data Warehouse

Many specialized database systems exist throughout the world that focuses on 2-D electrophoresis data, with SWISS-2DPAGE being one of the major sources [19]. However, these run in isolation and rarely cooperate with each other, and there is not a reliable way to allow comparisons to be made between data sets. This raises the problem of data exchange. Providing interoperability among these databases is important to the successful operation of an organization. Improvements in productivity will be gained if the systems can be integrated – that is, made to cooperate with each other – to support global applications accessing multiple databases. Hence, a common repository such as the Electrophoresis Data Repository (Fig. 5) proposed here for electrophoresis data would be ideal. This process model, described in the following sections, uses TWODML as a format for query results. Thus information from the repository can be shared with any

other system supporting TWODML. Also, XML tools may be used to make the query results human-viewable.

Data Source

One of the main objectives of the Electrophoresis Data Repository is to provide the community with detailed information about a given protein of interest, including qualitative and quantitative properties. This requires more comprehensive data for each of the data items and data groups. One way to achieve this is to encourage the electrophoresis community to deposit their experimental data, so this will facilitate the entry of electrophoresis data and associated information through a web based common repository. It is clear that having to supply such detail for every single parameter in an experiment data set can be highly onerous task for data submitters. If a public repository is to encourage submission of experimental data, the designers of such databases must strive to reduce the amount of manual labor required of submitters. Software tools accompanying electronic repositories must provide the equivalent of GeneBank's SEQUIN (http://www.ncbi.nlm.nih.gov/Sequib), BankIt (http://www.ncbi.nlm.nih.gov/BankIt), or Gene Expression Omnibus (GEO: http://www.ncbi.nlm.nih.gov/geo/). Nonetheless, if an electrophoresis data repository has sufficient data in enough categories then it will be possible to query the required data. Also, a parallel effort has to be established to gather electrophoresis data from the published articles and integrate into the repository. This knowledge component of a resource is usually held in scientific natural language as text. Extracting 2-D experimental information could also be done using Information retrieval (IR). IR is the field of computer science that deals with the processing of documents containing free

text, so that they can be rapidly retrieved based on keywords specified in a user's query [20]. Data derived form the public repository as well as from the mined data can be stored in a raw data repository. Raw data should be archived in a standard format, so as to ensure the data integrity, originality, and traceability.

Data Annotation/Validation

Electrophoresis data are complex and heterogeneous, that should be annotated with the community's understanding about the data. Each data point and its value need to be examined or validated for its correctness and completeness. The annotation of data elements also requires that all of the related data records within a file are consistent and properly integrated across the group of files. The first phase will be largely automated; annotation resources that will be routinely consulted to provide a complete range of updated, annotated information. Continuous annotation of electrophoresis data will improve the interoperability of cross-platform data sets. Since we are adopting TWODML, the annotation of electrophoresis data is also extendable to Distributed Annotation System (DAS). DAS is a client-server system in which a single client integrates information from multiple servers to retrieve and integrate dispersed proteomics data [21]. Since DAS adopts the integration through XML, well-annotated electrophoresis data can be integrated from various specialized resources through TWODML.

Data Management and Archiving

The volume of electrophoresis data is growing at a nearly exponential rate and this poses a problem in terms of data management, scalability, and performance. Building of the database structure is the first step towards the structured recording of electrophoresis data

in a relational database. This consists of precisely defined data fields and precisely defined relationships between them represented by links between the tables. Electrophoresis data are complex to model and there are many different types of data presenting numerous relationships. Data models are the logical structures used to represent a collection of entries and their underlying one-to-one, one-to-many, and many-to-many relationships. The main motivation for creating electrophoresis data models is usually to be able to implement them within database management systems, usually as a relational database management system (e.g., ORACLE, SQL Server, SYBSASE, MySQL, etc.). This electrophoresis data model corresponds to a way of organizing the pertinent values obtained on measurement of an electrophoresis experiment. This database is being modeled to handle the heterogeneous data from various external data sources. New types of proteomics data emerge regularly and this raises the need for updating the whole data semantics, and integrating the sources of information that were formally independent. Data analysis generates new data that also have to be modeled and integrated.

Query-able Web Interface

A customizable query interface should permit complex queries that include the most commonly required data for electrophoresis. The individual or grouped data should be able to be downloaded in TWODML format.

Data Distribution

Data are accessed intensively and exchanged very often by users on the Internet. Users can gather their data in TWODML format through a web interface that can be queried. This is useful for data exchange between other systems compatible with TWODML. On

the other hand, users might also want to view the data. In this case, the data must be presented in a user-friendly format. An obvious choice is HTML since it is supported by all Internet browsers. Another, perhaps more compelling format useful for human viewing is SVG (Scalable Vector Graphics) – a standard XML format for describing two-dimensional graphics. Unlike vanilla HTML, SVG drawings are dynamic, interactive, and may even be animated. HTML and/or SVG can be obtained from TWODML query results using XSLT.

A Proposed System

Standards are adopted by: 1) Single vendor controls a large enough portion of the market to make its product the market standards; 2) Community agrees on an available standard specification; 3) Group of volunteers representing interested parties work in an open process to create a standard; 4) Government agency such as the National Institute of Standards and Technology (NIST) coordinates the creation of consensus standards.

The National Institute of Standards and Technology (NIST) has an established record for the development and certification of validated and standardized databases. NIST is currently developing a program to meet some of the needs of the proteomics communities. As an exploratory model system for 2-D PAGE data standards, we experimented with the 2-D PAGE data of mitochondrial proteins [22]. Mitochondrion is an ideal target for global 2-D PAGE because of its manageable level of complexity.

Theoretical Model

As an initial attempt at standardization, implementation of a 2-D PAGE reference based upon the theoretical calculation of isoelectric point (pI) and molecular weight for each

mitochondrial protein (both mitochondrial and nuclear encoded) sequences has been carried out. A customizable interface was developed to permit complex queries that include the name of the protein, tissue, mitochondrial compartment, chromosome number where relevant, molecular weight range, pI range, and keywords (Fig. 6). The query results, along with the protein name, pI and molecular weight are presented as an intermediate selection, with the protein name linked to the theoretical virtual 2-D PAGE of that protein. The virtual 2-D PAGE shows the query protein's mobility on a virtual two-dimensional gel, based upon the isoelectric point and molecular weight as calculated from the protein sequence (Fig. 7). Each 2-D spot's protein name is linked to detailed information about the protein (Fig. 8). The information presented on the detail page includes: SwissProt information, description, cellular location, key words, tissue, cellular function, similarity with other proteins, gene name, synonyms, chromosomal location, and protein sequence information such as the amino acid length, theoretical pI, and theoretical molecular weight. External database links are also presented on this page. The protein sequence can be highlighted using a mouse-over option that provides annotation such as the mitochondrial localization signal, variant information, and other protein sequence details. The protein sequence of interest can also be used to search the journal references, the SwissProt site for related proteins, or against the Protein Data Bank sequence for the related 3-D structures.

Experimental Model

We plan to extend the 2-D PAGE option of our theoretical model into an experimental model. In order to be able to search a 2-D PAGE database with selectable query options for an individual protein, it is necessary to have comprehensive data for each of the

individual categories. Currently, we are experimenting to mine the 2-D PAGE data from the published literature and incorporate into our model data warehouse. The issue of effectively describing the data elements for spot intensity will be critical and should thus be carefully considered. The spot intensity, direct image file data, should be considered separately from protein concentration reporting, derived data from the image intensity. Physical calibration standards will probably be required to be captured in order to permit one to associate protein concentration with image intensity. The issue of recording spot intensity data has been worked out in other disciplines, such as the much more complicated x-ray crystallography case, where image plates or detectors capture x-ray diffraction reflections. Applications of these or similar algorithms would seem prudent. We are also in the process of exploring physical 2-D PAGE standards and associated algorithms to dynamically detect and measure 2-D spots from diverse 2-D gels. Our mitochondrial 2-D PAGE information is available from the NIST mitochondrial protein web site at: http://bioinfo.nist.gov:8080/examples/servlets/index.html

A community effort towards a similar goal for 2-D PAGE meta data is urgently needed. Individual research consortia, such as Proteomics Standards Initiative (PSI), Human Proteome Organization (HUPO), or the electrophoresis society are in a good position to explore informatics approaches that will work within a consortium, but the time is ripe for the formation of ad hoc groups at various proteomics meetings, and it is hoped that this article acts as the catalyst for such collaboration.

“Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?” – T. S. Elliott

Acknowledgements

This work was supported in part by an Exploratory Research Grant (VR) from the Chemical Sciences and Technology Laboratory (CSTL), National Institute of Standards and Technology (NIST). We are grateful to Drs. Anne Plant (NIST), John Kasianowicz (NIST), Richard B. Johnston, Jr. (University of Colorado) for intellectual contributions to this manuscript. We thank Sundari Ravi for her technical assistance.

Disclaimer

The contents of this article do not necessarily reflect the views or policies of National Institute of Standards and Technology (NIST). Any mention of commercial products within this article is for information only and does not imply recommendation or endorsement by NIST. The World Wide Web pages are provided as a public service by NIST. With the exception of material marked as copyright, information presented on these pages is considered public information and may be distributed or copied. Use of appropriate byline/photo/image credits is requested.

References

[1] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al., *Science*, 2001, *291*, 1304–1351.

[2] Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gestetland, R., Walters, L., *Science*, 1998, *282*, 682–689.

[3] Collins, F.S., McKusick, V.A., *J. Am. Med. Assoc.* 2001, *285*, 540–544.

[4] Molloy, M.P., *Nat. Biotechnol.* 2003, *21*, 597.

[5] Frank, R., Hargreaves, R., *Nat. Rev. Drug Discov.* 2003, *2*, 566-580.

[6] Petricoin, E.F., Zoon, K.C., Kohn, E.C., Barrett, J.C., Liotta, L.A., *Nat. Rev. Drug Discov.* 2002, *1*, 683-695.

[7] Renfrey, S., Featherstone, J., *Nat. Rev. Drug Discov.* 2002, *1*, 175-176.

[8] Verrills, N.M., Kavallaris, M., *Curr. Opin. Mol. Ther.* 2003, *5*, 258-265.

[9] Srinivas, P.R., Verma, M., Zhao, Y., Srivastava, S., *Clin. Chem.* 2002, *48*, 1160-1169.

[10] Goldstein, D., *Manag. Care Interface.* 2003, *16*, 68-69.

[11] Stupka, E., *Curr. Opin. Mol. Ther.* 2002, *4*, 265-274.

[12] Campagne, F., *Nature*, 2002, *418*, 125.

[13] Jones, J., Preston, H., *Top. Health Inf. Manage.* 2000, *21*, 45-54.

[14] Achard, F., Vaysseix, G., Barillot, E., *Bioinformatics*, 2001, *17*, 115-125.

[15] Mackenzie, D., *Science*, 1998, *280*, 1840-1841.

[16] World Wide Web Consortium, Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation, 6 October 2000. http://www.w3.org/TR/REC-xml

[17] World Wide Web Consortium, XML Schema Part 1: Structures, W3C Recommendation, 2 May 2001, http://www.w3.org/TR/xmlschema-1/

[18] Organization for the Advancement of Structured Information Systems (OASIS), RELAX NG Specification, Committee Specification, 3 December 2001, http://relaxng.org/spec.html

[19] Hoogland, C., Sanchez, J.C., Tonella, L., Binz, P.A., Bairoch, A., Hochstrasser, D.F., Appel, R.D., *Nucleic Acids Res.* 2000, 28, 286-288.

[20] Nadkarni, P.M., *Pharmacogenomics J.* 2002, *2*, 96-102.

[21] Stein, L.D., *Nat. Rev. Genet.* 2003, *4*, 337-345.

[22] Verma, M., Kagan, J., Sidransky, D., Srivastava, S., *Nature Rev. Cancer*, 2003, *3*, 789-795.

Table 1: Biological Data Exchange Formats

Format	PROS	CONS
Flat files	The simplicity of the flat file, which does lend itself to simple tools that are available to all, is in great part responsible for its popularity. Flat files, based on field/value, are commonly used to represent biological data.	Lacks referencing typed values vocabulary control constraints, among other issues. Often fields are ambiguous and their content is contextual.
Abstract Syntax Notation One (ASN.1)	ASN.1 files convey the description of their structure and offer some flexibility; the client side does not necessarily need to know in advance the structure of the data. Based on that single, common format, a number of human-readable formats and tools are produced, such as those used by Entrez, GenBank, and the BLAST databases.	ASN.1 software tools do not scale well for very large data sets, and there is no support for queries. Workers on ASN.1 Standards recognized several years ago that there was a requirement (from users of ASN.1) to have an XML representation of the information structures defined by an ASN.1 specification.
Common Object Request Broker Architecture (CORBA)	CORBA provides platform-independent programming interfaces and models for portable distributed object-oriented computing applications. Its independence from programming languages, computing platforms and network protocols provides a solution for developing new applications for querying and distributing biological data, which can also be integrated into existing systems.	Despite the benefits of developing a CORBA environment, it's a heavy task that requires highly skilled computer scientists and expensive software. Further, CORBA messages are blocked by most firewalls, making CORBA an impractical option for Internet-accessible systems.
Java Remote Method Invocation (RMI)	RMI enables the programmer to create distributed Java technology-based to Java technology-based applications, in which	Like CORBA, its scalability depends greatly on the network bandwidth.

	the methods of remote Java objects can be invoked from other Java virtual machines, possibly on different hosts.	
Object-oriented Database Management System (ODBMS)	ODBMS extend the object programming language with transparently persistent data, concurrency control, data recovery, associative queries, and other database capabilities. Offers rich data model well suited to biology.	High learning curve to fully understand the technology

Table 2: Scientific Markup Languages

Markup Language	**Purpose**	**URL**
Chemical Markup Language (CML)	Exchange of chemical information	http://www.xml-cml.org
Mathematical Markup Language (MathML)	Exchange of mathematical formula.	http://www.w3.org/Math
Bioinformatic Sequence Markup Language (BSML)	Exchange of DNA, RNA, protein sequences and their graphic properties	http://www.sbw-sbml.org/index.html
BIOpolymer Markup Language (BIOML)	Expression of complex annotation for protein and nucleotide sequence information.	http://www.bioml.com/BIOML
The taxonomical markup language	Exchange of taxonomic relationships between organisms.	http://www.albany.edu/~gilmr/pubxml
Genome Annotation Markup Elements (GAME)	Annotation of biosequence features	http://xml.coverpages.org/game.html
BlastXML	Model NCBI Blast output.	http://doc.bioperl.org/releases/bioperl-1.2/Bio/SearchIO/blastxml.html
Ontology Mark-up Language/Conceptual Knowledge Markup Language (OML/CKML)	Representation of biological knowledge and specifically functional genomic relationships.	http://smi-web.stanford.edu/projects/bio-ontology
Multiple Sequence Alignments Markup Language (MSAML)	Description of multiple sequence alignments (amino acids and nucleic acid sequences)	http://xml.coverpages.org/msaml.html
Systems Biology Markup Language (SBML)	Representation and modeling of the information components in the system biology	http://www.cds.caltech.edu/erato/sbml/docs
Gene Expression Mark-up Language (GEML)	Exchange of gene expression data, Gene Expression Mark-up Language	http://www.oasis-open.org/cover/geml.html
GeneX Gene Expression Markup Language (GeneXML)	Representation of the Gene Expression Databases datasets	http://xml.coverpages.org/geneXML.html
Microarray Markup Language (MAML)	Integration of microarray data	http://xml.coverpages.org/maml.html

Protein Markup Language (ProML)	Exchange of protein sequences, structures, and families based data	http://www.bioinfo.de/isb/gcb01/talks/hanisch/main.html
RNA Markup Language (RNAML)	Exchange of RNA information	http://www-lbit.iro.umontreal.ca/rnaml/

Figure 1

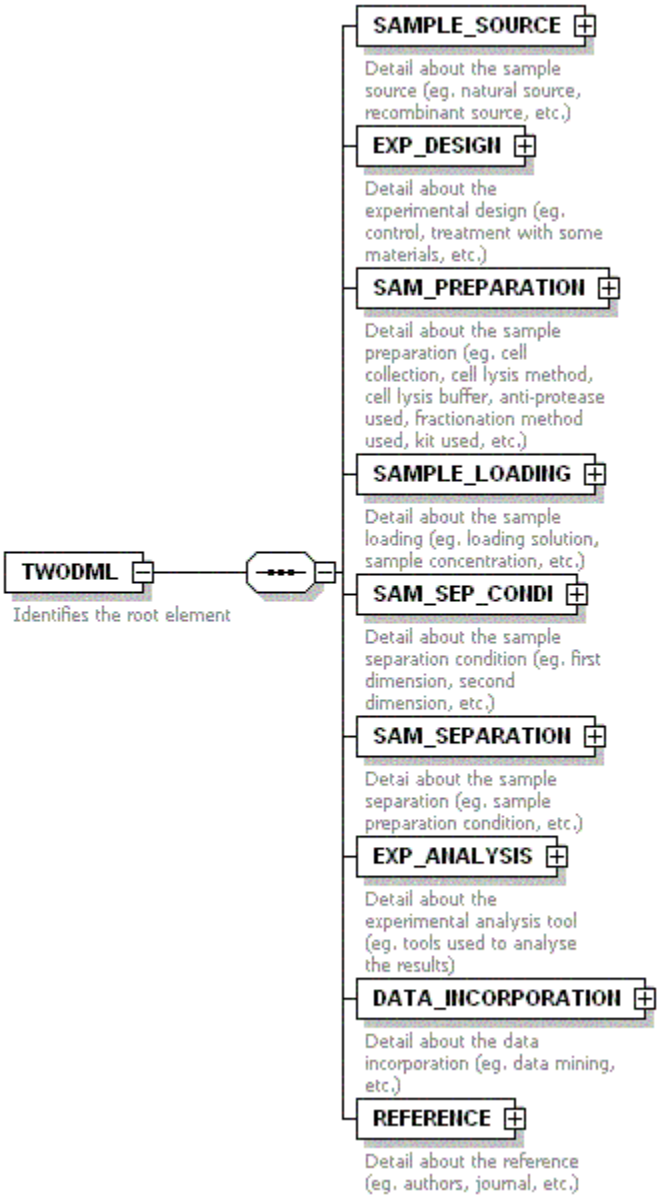
TWODML
Identifies the root element
SAMPLE_SOURCE
Detail about the sample source (eg. natural source, recombinant source, etc.)
EXP_DESIGN
Detail about the experimental design (eg. control, treatment with some materials, etc.)
SAM_PREPARATION
Detail about the sample preparation (eg. cell collection, cell lysis method, cell lysis buffer, anti-protease used, fractionation method used, kit used, etc.)
SAMPLE_LOADING
Detail about the sample loading (eg. loading solution, sample concentration, etc.)
SAM_SEP_CONDI
Detail about the sample separation condition (eg. first dimension, second dimension, etc.)
SAM_SEPARATION
Detai about the sample separation (eg. sample preparation condition, etc.)
EXP_ANALYSIS
Detail about the experimental analysis tool (eg. tools used to analyse the results)
DATA_INCORPORATION
Detail about the data incorporation (eg. data mining, etc.)
REFERENCE
Detail about the reference (eg. authors, journal, etc.)

Figure 2

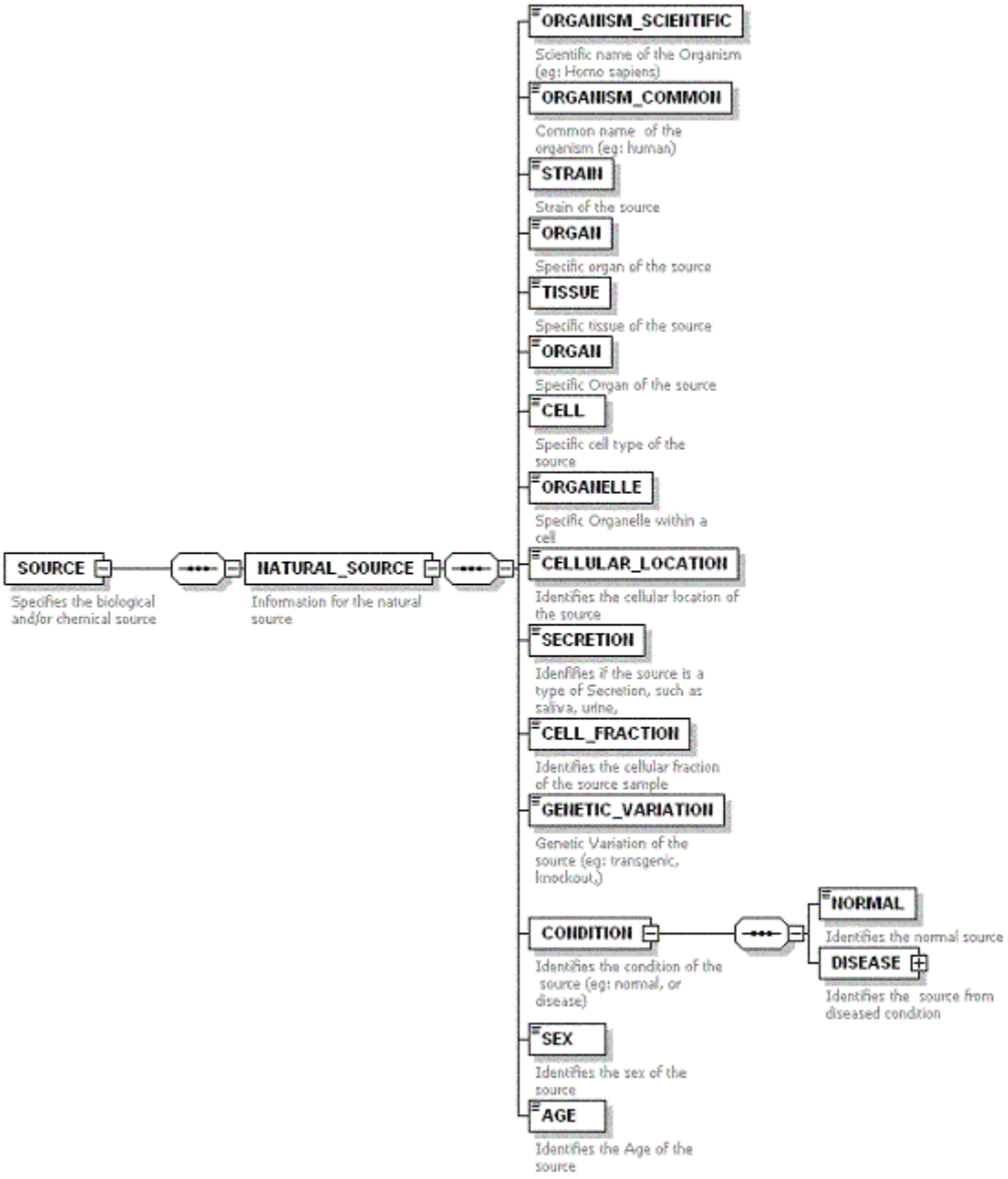
SOURCE
Specifies the biological and/or chemical source
NATURAL_SOURCE
Information for the natural source
ORGANISM_SCIENTIFIC
Scientific name of the Organism (eg: Homo sapiens)
ORGANISM_COMMON
Common name of the organism (eg: human)
STRAIN
Strain of the source
ORGAN
Specific organ of the source
TISSUE
Specific tissue of the source
ORGAN
Specific Organ of the source
CELL
Specific cell type of the source
ORGANELLE
Specific Organelle within a cell
CELLULAR_LOCATION
Identifies the cellular location of the source
SECRETION
Identifies if the source is a type of Secretion, such as saliva, urine,
CELL_FRACTION
Identifies the cellular fraction of the source sample
GENETIC_VARIATION
Genetic Variation of the source (eg: transgenic, knockout,)
CONDITION
Identifies the condition of the source (eg: normal, or disease)
NORMAL
Identifies the normal source
DISEASE
Identifies the source from diseased condition
SEX
Identifies the sex of the source
AGE
Identifies the Age of the source

Figure 3

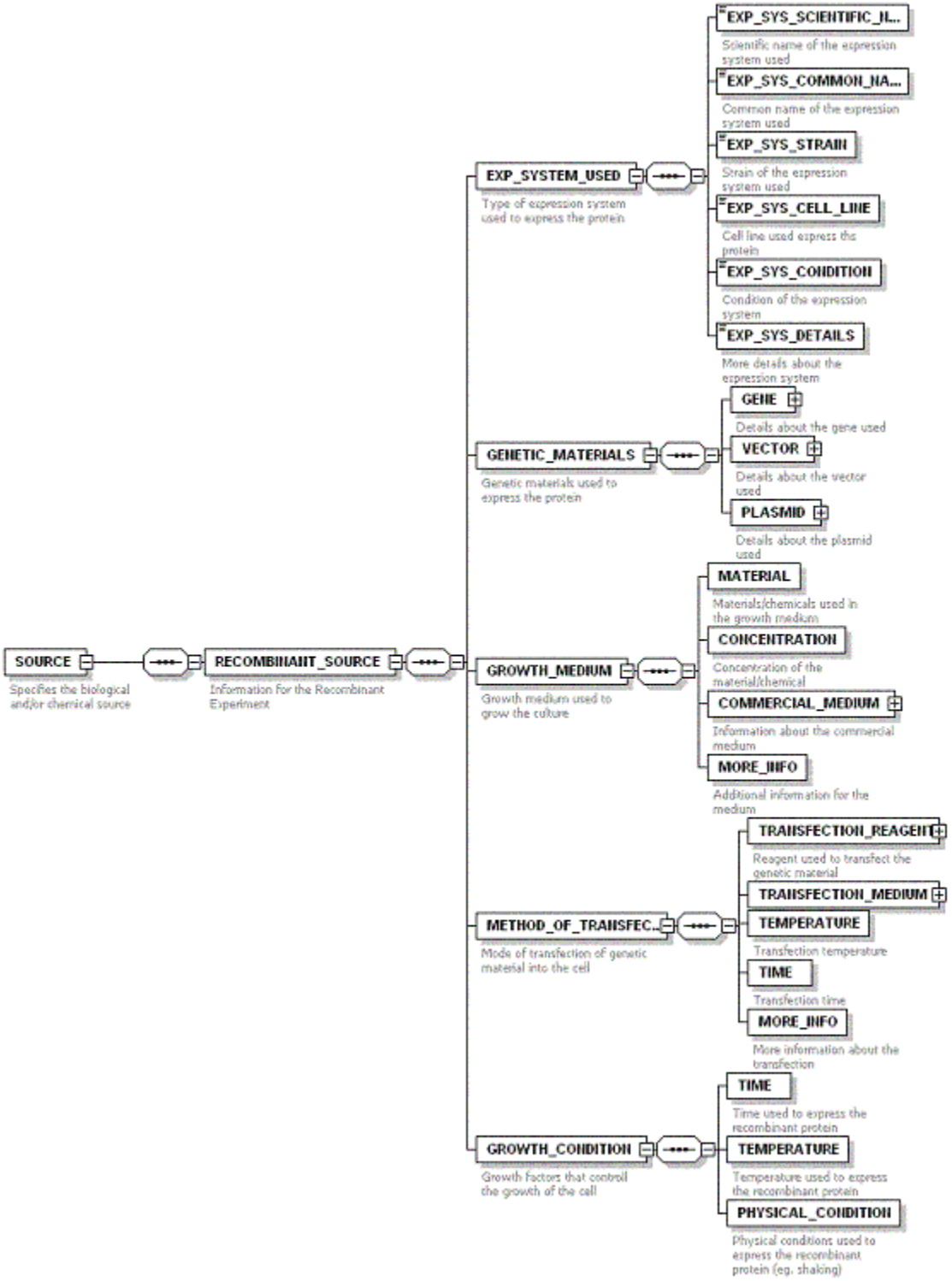
SOURCE
Specifies the biological and/or chemical source
RECOMBINANT_SOURCE
Information for the Recombinant Experiment
EXP_SYSTEM_USED
Type of expression system used to express the protein
EXP_SYS_SCIENTIFIC_N...
Scientific name of the expression system used
EXP_SYS_COMMON_NA...
Common name of the expression system used
EXP_SYS_STRAIN
Strain of the expression system used
EXP_SYS_CELL_LINE
Cell line used express the protein
EXP_SYS_CONDITION
Condition of the expression system
EXP_SYS_DETAILS
More details about the expression system
GENETIC_MATERIALS
Genetic materials used to express the protein
GENE
Details about the gene used
VECTOR
Details about the vector used
PLASMID
Details about the plasmid used
GROWTH_MEDIUM
Growth medium used to grow the culture
MATERIAL
Materials/chemicals used in the growth medium
CONCENTRATION
Concentration of the material/chemical
COMMERCIAL_MEDIUM
Information about the commercial medium
MORE_INFO
Additional information for the medium
METHOD_OF_TRANSFEC...
Mode of transfection of genetic material into the cell
TRANSFECTION_REAGENT
Reagent used to transfect the genetic material
TRANSFECTION_MEDIUM
TEMPERATURE
Transfection temperature
TIME
Transfection time
MORE_INFO
More information about the transfection
GROWTH_CONDITION
Growth factors that control the growth of the cell
TIME
Time used to express the recombinant protein
TEMPERATURE
Temperature used to express the recombinant protein
PHYSICAL_CONDITION
Physical conditions used to express the recombinant protein (eg. shaking)

Figure 4

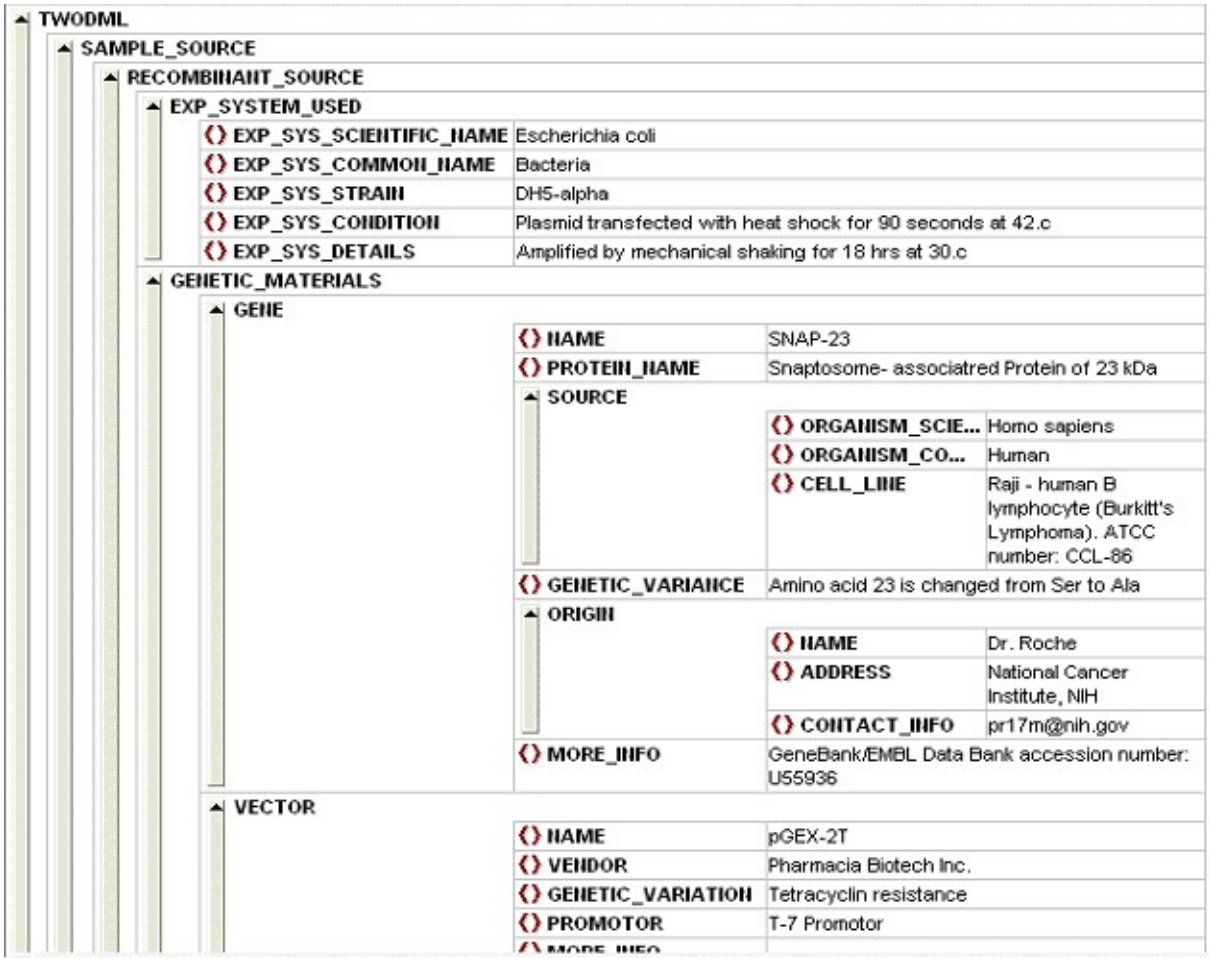
TWODML
SAMPLE_SOURCE
RECOMBINANT_SOURCE
EXP_SYSTEM_USED
EXP_SYS_SCIENTIFIC_NAME
Escherichia coli
EXP_SYS_COMMON_NAME
Bacteria
EXP_SYS_STRAIN
DH5-alpha
EXP_SYS_CONDITION
Plasmid transfected with heat shock for 90 seconds at 42.c
EXP_SYS_DETAILS
Amplified by mechanical shaking for 18 hrs at 30.c
GENETIC_MATERIALS
GENE
NAME
SNAP-23
PROTEIN_NAME
Snaptosome- associatred Protein of 23 kDa
SOURCE
ORGANISM_SCIE...
Homo sapiens
ORGANISM_CO...
Human
CELL_LINE
Raji - human B lymphocyte (Burkitt's Lymphoma). ATCC number: CCL-86
GENETIC_VARIANCE
Amino acid 23 is changed from Ser to Ala
ORIGIN
NAME
Dr. Roche
ADDRESS
National Cancer Institute, NIH
CONTACT_INFO
pr17m@nih.gov
MORE_INFO
GeneBank/EMBL Data Bank accession number: U55936
VECTOR
NAME
pGEX-2T
VENDOR
Pharmacia Biotech Inc.
GENETIC_VARIATION
Tetracyclin resistance
PROMOTOR
T-7 Promotor

Figure 5

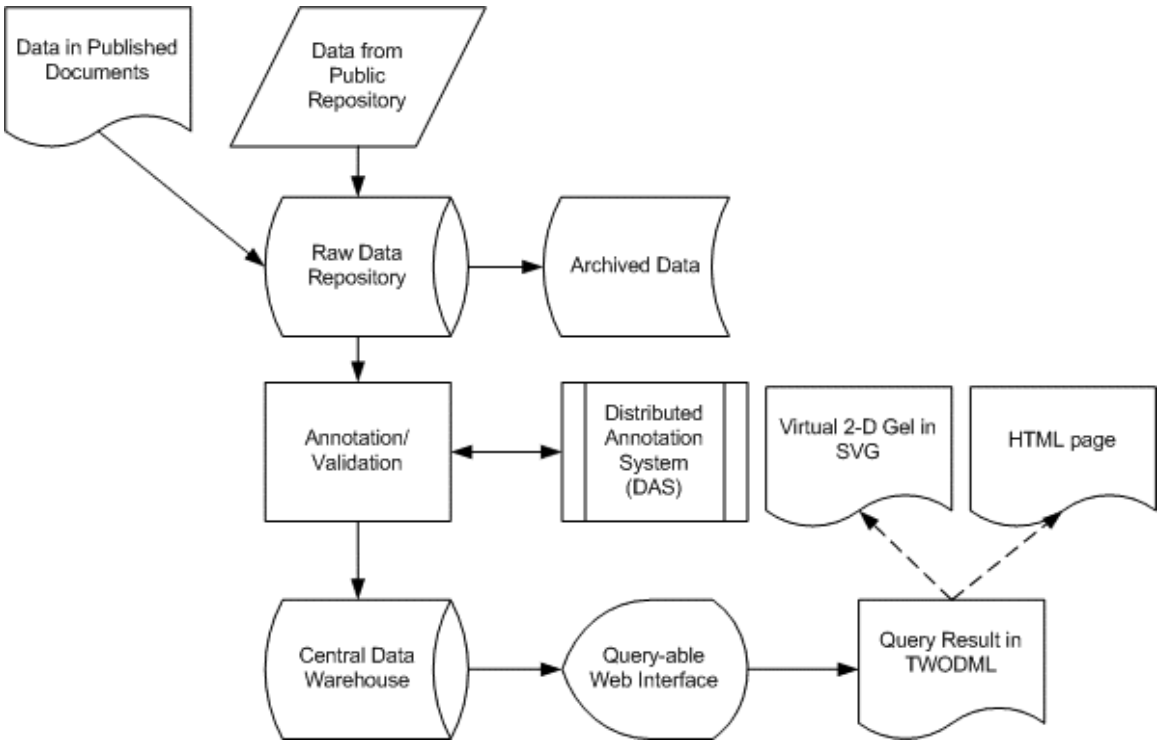
Data in Published Documents
Data from Public Repository
Raw Data Repository
Archived Data
Annotation/ Validation
Distributed Annotation System (DAS)
Virtual 2-D Gel in SVG
HTML page
Central Data Warehouse
Query-able Web Interface
Query Result in TWODML

Figure 6

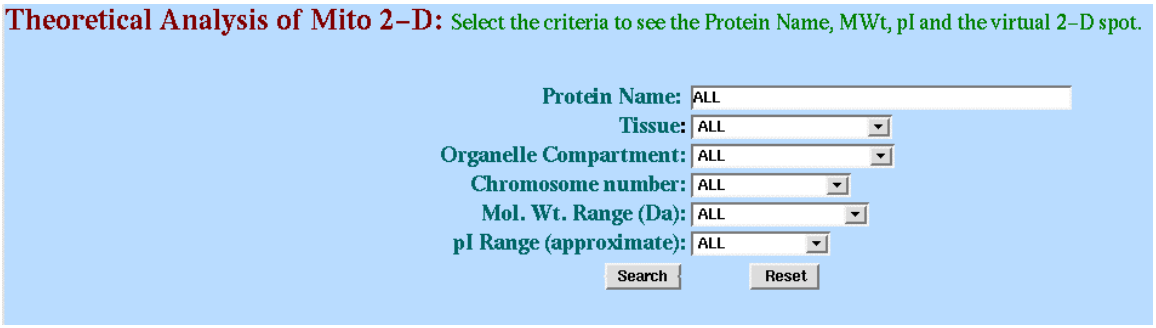
Theoretical Analysis of Mito 2–D: Select the criteria to see the Protein Name, MWt, pI and the virtual 2–D spot.
Protein Name: ALL
Tissue: ALL
Organelle Compartment: ALL
Chromosome number: ALL
Mol. Wt. Range (Da): ALL
pI Range (approximate): ALL
Search
Reset

Figure 7

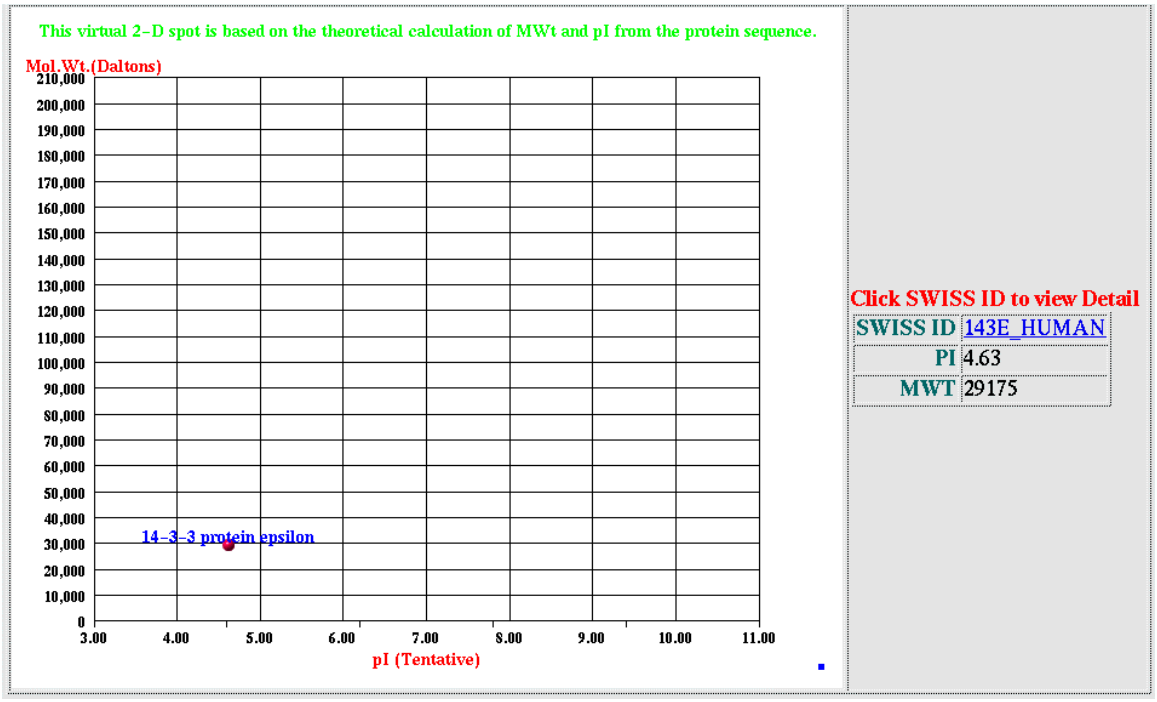
This virtual 2-D spot is based on the theoretical calculation of MWt and pI from the protein sequence.
Mol.Wt.(Daltons)
210,000
200,000
190,000
180,000
170,000
160,000
150,000
140,000
130,000
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
14-3-3 protein epsilon
3.00
4.00
5.00
6.00
7.00
8.00
9.00
10.00
11.00
pI (Tentative)
Click SWISS ID to view Detail
SWISS ID 143E_HUMAN
PI 4.63
MWT 29175

Figure 8

Detail Information for 143E_HUMAN

SwissProt ID	143E_HUMAN
Description	14–3–3 protein epsilon (Mitochondrial import stimulation factor Lsubunit) (Protein kinase C inhibitor protein–1) (KCIP–1) (14–3–3E)
Location	CYTOPLASMIC
Key Words	Brain; Neurone; Acetylation; Multigene family
Tissue	Heart;Liver;Placenta
Function	ACTIVATES TYROSINE AND TRYPTOPHAN HYDROXYLASES IN THE PRESENCE OF CA(2+)/CALMODULIN–DEPENDENT PROTEIN KINASE II, AND STRONGLY ACTIVATES PROTEIN KINASE C. IS PROBABLY A MULTIFUNCTIONAL REGULATOR OF THE CELL SIGNALING PROCESSES MEDIATED BY BOTH KINASES
Similarity	BELONGS TO THE 14–3–3 FAMILY
Gene	YWHAE
Aliases	14–3–3 epsilon
OMIM ID	605066
RefSeq ID	006761
Locus Link ID	7531
Chromosomal Location	17p13
Gene Database ID	217083
PubMed ID	9371399

Sequence Information for 143E_HUMAN

pI (tentative)	4.6
Amino Acid Length	255
Molecular WT(Daltons)	29175

(Mouseover on the colored sequence for Detail)

M D D R E D L V Y Q A K L A E Q A E R Y D E M V E S M K K V A G M D V E L T V E E R N L L S V A Y K N V I G A R R A S W R I I S S I E Q
K E E N K G G E D K L K M I R E Y R Q M V E T E L K L I C C D I L D V L D K H L I P A A N T G E S K V F Y Y K M K G D Y H R Y L A E F A
T G N D R K E A A E N S L V A Y K A A S D I A M T E L P P T H P I R L G L A L N F S V F Y Y E I L N S P D R A C R L A K A A F D D A I A E L
D T L S E E S Y K D S T L I M Q L L R D N L T L W T S D M Q G D G E E Q N K E A L Q D V E D E N Q

Search this Sequence against SwissProt | Search this Sequence against Protein Data Bank(PDB)

View References

View the above info in Extensible Markup Language(XML):

XML File | XML DTD File

Figure Legends

Figure 1: Schematic diagram for the TWODML

Figure 2: Schematic diagram for the natural source

Figure 3: Schematic diagram for the recombinant source

Figure 4: A part of recombinant source XML model data

Figure 5: Process model for 2-D data warehouse

Figure 6: Query-able web interface for mitochondrial 2-D (theoretical) data

Figure 7: Virtual 2-D gel image for 14-3-3 protein

Figure 8: An example of the detailed information page from a search for protein 14-3-3.