RI 02017, 10 January 2002 Computer Science

IBM Research Report

A Common Data Representation for Organizing and Managing Annotations of Biochip Expression Data

Sudeshna Adak, Biplav Srivastava, Pankaj Kankar, Manish P Kurhekar
IBM Research Division
IBM India Research Lab
Block I, I.I.T. Campus, Hauz Khas
New Delhi - 110016, India.

IBM Research Division
Almaden - Austin - Beijing - Delhi - Haifa - T.J. Watson - Tokyo - Zurich

A Common Data Representation for Organizing and Managing Annotations of Biochip Expression Data

Sudeshna Adak, Biplav Srivastava, Pankaj Kankar, Manish P Kurhekar

IBM India Research Lab
Block 1, Indian Institute of Technology
Hauz Khas
New Delhi, India 110016

Abstract

This paper describes a systematic method of representation for data related to "biochip" experiments. The data is either a) gene or protein expression data as obtained from biochip experiments or b) information related to genes and proteins of interest in the biochip experiment, which has been extracted from heterogeneous data sources. The goal is to synthesize information from disparate types of biological data. The objective of this paper is to define a common representation for the disparate types of biological data in a manner that allows modelling, querying, and annotating of the biochip experimental data.

BACKGROUND

Biochips:

The term biochip refers to miniaturized microfluidic systems for performing massively parallel biochemical assays. A biochip has a matrix of "probes" affixed to glass, silicon, or nylon substrates. In a biochip experiment, the probes react with a biological sample that has been labelled with a fluorescent dye. The reaction simultaneously performs thousands of biochemical assays. The results of the assays are measured by fluorescence intensity. Microarray is a term also used to refer to biochips.

In this paper, the term "biochip" will include

1. DNA chips or DNA microarrays used to study mRNA expression (probe used is cDNA or oligonucleotides);
2. Protein chips, protein arrays and protein microarrays, antibody microarrays used to study protein expression and regulation and protein-protein interactions (probe used is an antibody) ;
3. PNA (peptide nucleic acid) arrays which use PNAs as probes;
4. Genome arrays or genome chips in which the genome of entire organisms is placed on a single chip and meant to monitor the expression levels in the genes/proteins of entire genomes;
5. New and emerging technologies of microarray fabrication that use small biomolecules.

The term "biochip" will also include (depending on the type of substrate used): high-density arrays (silicon substrate), medium-density spotted arrays (glass slide used as substrate), medium to low-density spotted arrays (nylon substrate) and new/rediscovered substrates for use in microarray fabrication.

Real world applications of biochips/microarrays are expected to impact the drug discovery process as well as clinical practice over the next 10 years. A few representative examples of such applications include:

1. Target Gene Discovery: Biochips are enabling scientists in the lab to study thousands of gene/protein expression profiles. By comparing the expression profiles of diseased cells versus normal cells and determining the differentially expressed genes/proteins, it is possible to discover sets of genes or proteins that play key roles in diseases (Debouck and Goodfellow, 1999) . These key sets of genes or proteins can be very useful as initial targets for development of therapeutic drugs.
2. SNP chips: Microarrays can be used for genome-wide detection of single nucleotide polymprphisms (Chakravarti, 1999).
3. Diagnostic Chips: Biochips can be customized to detect a disease-specific pattern of gene/protein chip expression from tissue or peripheral blood samples. Such tailor-made biochips will be used as diagnostic kits in case of diseases which can be accurately identified by this technique (Tamayo et al, 1999).
4. Pharmacology and Toxicology: The use of microarrays in screening a large number of compounds for efficacy and toxicity in cell cultures has already been established (Scherf et al, 2000). Biochips can enable pharmaceutical researchers to screen drugs more effectively in determining target compounds for the more expensive and time-consuming clinical trials.
5. Personalized Medicine: Biochips are able to provide information about drug efficacy and how the human body interacts with drugs at the molecular level (Scherf et al, 2000). This type of information is necessary to pave the way for personalized medicine, where treatment of disease can be tailored to the genetic profile of individuals.

Given the nature of the impact of biochips in the pharmaceutical and agricultural industry, a common data representation that facilitates modelling, querying, and annotating of biochip data is clearly of practical significance.

DESCRIPTION OF THE PROBLEM

The need for a Common Data Representation:

A complete and meaningful analysis of biochip experimental data involves:

1. Statistical/Exploratory analyses of the gene/protein expression data as obtained from a series of biochip experiments. The standard techniques in current use for such analyses are two-way hierarchical clustering (Eisen et al, 1998) and self organizing maps (Tamayo et al, 1999).
2. Annotation of the expression data with the results of any analyses and relevant information derived from heterogeneous databases.
3. Knowledge discovery through querying, analyzing, data mining and visualization of the experimental and annotation information. The term knowledge discovery as used in this

context means synthesis of relevant biological information derived from biochip experiments and disparate data sources. Examples include

- Discovery of functional relationships between similarly expressed genes (or between similarly expressed proteins);
- Discovery of biological mechanisms by which genes are regulated;
- Relationships of genes and proteins to phenotypes (observed states) like disease, environmental conditions, drug dosage, etc.

An end-to-end Bioinformatics solution for biochips provide the framework and tools for implementing all of the above three tasks. A common data representation is a vital necessity in creating such a solution due to the following reasons.

1. The annotation information is available through a multitude of diverse data sources: A rich domain knowledge of possible choice of data sources is essential. A detailed knowledge of the capabilities and schema of relevant biological databases is required.
2. The annotation information is heterogeneous: The relevant biological databases are heterogeneous - differences in the type of content, in the data format used, in their access interface, and in the database management system. Thus, the annotation information derived from the biological databases are also heterogeneous.
3. Integration of the annotation information is required as a first step to data mining: The "knowledge discovery" segment involves querying, data mining, and visualization of these annotations. Given the heterogeneous nature of the data sources, common representation of the annotation and experimental data is imperative.

Desired features of the Common Data Representation:

There are many possible configurations of a common representation. However, some of the main requirements for a "good" common representation are:

1. It should be able to capture the relevant biological information.
   Examples of biological information relevant to the task of annotation and mining include DNA and protein sequence data, chromosomal location, protein structure, protein domains and protein families, pathways, disease information, and medline references.
2. It should be able to capture the relationship between data items of interest.
   Since it is of interest to combine information from various biological sources, it is necessary to define a representation that will capture the relationship between the "probes" in the biochip and the data items in the biological databases. For example, in a DNA microarray experiment which uses cDNA probes from cDNA clone libraries, it is necessary to associate the clone identifier with unigene cluster identifiers.
3. It should be non-redundant.
   A particular data item may be available from multiple databases. For example, protein sequence data is available from both Genbank and Swiss-Prot. It is desirable that the common representation be non-redundant and capture this information at a single location.
4. It should be in a form that facilitates browsing, querying, retrieval, interaction and efficient manipulation.
   The objective of the common representation is to facilitate modelling, querying, and annotation of biochip experiments. In order to achieve this objective, it is required that

the format of the common representation be particularly suited to browsing, querying, retrieval, interaction and efficient manipulation.

5. It should be compact.
   The common representation gives the user a unified view of disparate biological data. This unified view should provide a sophisticated and compact representation of the data hierarchy that might be returned by a complex query.
6. It should be scalable.
   It should be easily possible to add information from new databases and databases not already incorporated in the system.
7. It should be flexible.
   The schema for the common representation should not be fixed or rigid. This is essential as data items as well as entire databases can get consolidated, reformatted, or are discarded. In such cases, a flexible common representation is necessary as it can be easily adapted to changes.

XML SCHEMA FOR MICROARRAYS

XML:
XML, a standard maintained by the World Wide Web Consortium (W3C), defines a syntax that can be used to create new markup languages. XML was chosen as the desired format for the common data representation in the context of biochips due to several compelling reasons:

1. XML provides the necessary syntax for accessing complex data sets over multiple platforms;
2. XML is flexible instead of fixed;
3. XML is hardware and software independent;
4. Information-based instead of layout-based;
5. Easy to read and work with: Text file-based and human readable;
6. Highly compatible with the web.
7. GEML™ (Gene Expression Markup Language), is an XML based format developed by Rosetta Inpharmatics and others to describe, store, and exchange expression data.

Some of the desired features of a “good” common representation mentioned while describing the problem is that the representation be scalable, flexible, and be suited to browsing and querying. By using XML, it is ensured that common representation will have these features.

We describe some of the extensible markup languages (XML) that have been proposed previously for biochip experiments.

1. MAML (http://www.oasis-open.org/cover/maml.html)
   The Microarray Markup Language (MAML) was proposed in order to facilitate the establishing of gene expression data repositories, comparability of gene expression data from different sources and inter-operability of different gene expression databases and data analysis software. The proposal was put forward by the microarray gene expression database group, MGED (http://www.mged.org/). The proposed framework was for describing information about a DNA-array experiment and a data format for communicating this information. The information included details about: (1) Experimental design: the set of the hybridization experiments as a whole; (2) Array design: each array

used and each element (spot) on the array; (3) Samples: samples used, the extract preparation and labeling; (4) Hybridizations: procedures and parameters; (5) Measurements: images, quantitation, specifications; (6) Controls: types, values, specifications. MAML was based on the Extensible Markup Language XML. MAML was independent of the particular experimental platform and provided a framework for describing experiments done on all types of DNA-arrays. The format allowed representation of raw and processed microarray data. The format was compatible with the definition of the 'minimum information about a microarray experiment' (MIAME) proposed by the MGED group, see http://www.mged.org/.

2. GeneXML (http://www.ncgr.org/genex/genexml.html)
   The GeneX Gene Expression Markup Language (GeneXML) began with the idea that the underlying communication channel would be a mechanism that would allow the complex annotations and protocols to be transmitted along with the numeric data in a way that would not require the explicit association of many files and certainly not by hand-editing different formats into a coherent whole. This project was started by the National Center for Genome Research (NCGR) in 1998 when there were no existing gene expression markup languages. NCGR now refers to its markup language as GeneXML, and is concentrating on increasing the utility and scalability of the underlying data model and moving towards full support of MAML. NCGR will also attempt to provide more compact and sophisticated representations that better represents the data hierarchy that might be returned from a complex query. As complete experiment sets encoded in XML format can be quite large, GeneXML breaks up its Document Type Definition (DTD) into logical components that can be downloaded separately. These include (1) GeneXML.dtd, which is the main GeneXML data structure that contains the experimental meta-data; (2) usf.dtd, which contains information on the immobilized sequence; (3) als.dtd, which handles the array layout information as it relates to the spots; (4) ams.dtd, where the actual measurement data is stored.

3. GEML (http://www.geml.org)
   Rosetta Inpharmatics' Gene Expression Markup Language, GEML™ has been adopted relatively widely by the industry. It provides a uniform syntax for storing and exchanging gene expression data. The term "reporter" is used to designate a row of the data matrix and the term "profile" is used to describe a single biochip experiment. In particular, the document type definition file GEMLProfile.dtd (available at http://www.geml.org/dtds.htm) is used to store and handle the numeric expression data from multiple biochip experiments in biochips.

SUMMARY

The aim of the paper are to provide a framework for modelling, querying, and annotating gene and protein expression data as obtained from biochips. This is achieved through the following objectives:

1. Define a system for integrating biochip experimental data with annotation information from heterogeneous, distributed databases;
2. Define a system for integrating biochip experimental data with annotation information from heterogeneous applications;

3. Define a common data representation framework for storing and querying of the biochip experimental data coupled with the annotation information from 1 and 2 above.
4. Define a unified view through the common representation which is compact, irredundant, flexible, and shows the data hierarchy.
5. Define an XML based markup language, ***eXpressML***, for the common data representation.

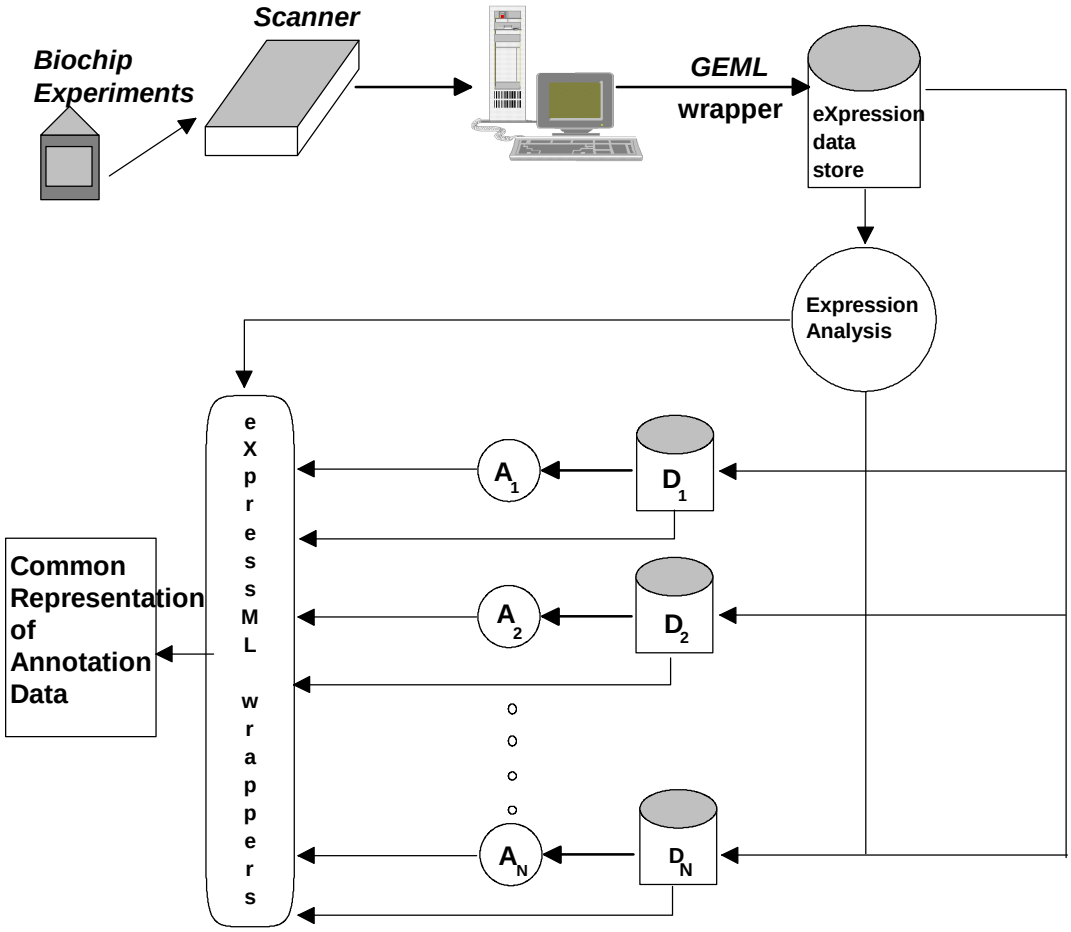

Figure 1: Schema for Creating and Storing Common Representation of Annotation Information Associated with Biochip Experiments

Standard model for expression data from biochip experiments: (See Figure 2 below)

The data from a series or set of biochip experiments is typically represented as a matrix of numbers, in which each row corresponds to a "spot" on the biochip that has been loaded with the biological "probe" and each column corresponds to a measurement from a single biochip experiment. (Multiple columns can be used to for a single experiment when storing different types of measurements such as red and green fluorescence intensities, background intensities, and calculated intensity ratios. The minimum data requirement for an experiment or profile is a single column containing the intensity ratio.)

We shall make use of the widely accepted Gene Expression Markup Language (GEML). The elements in the GEML files and how they are nested are shown below. Detailed examples and definition of the tags are given in http://www.geml.org.

<project>
<profile>

<hyb>
<channel_info>
<reporter>
<feature>
<channel>
<signal>
<background>
<intensity ratio>
<position>

Common representation of annotation data for biochip experiments:

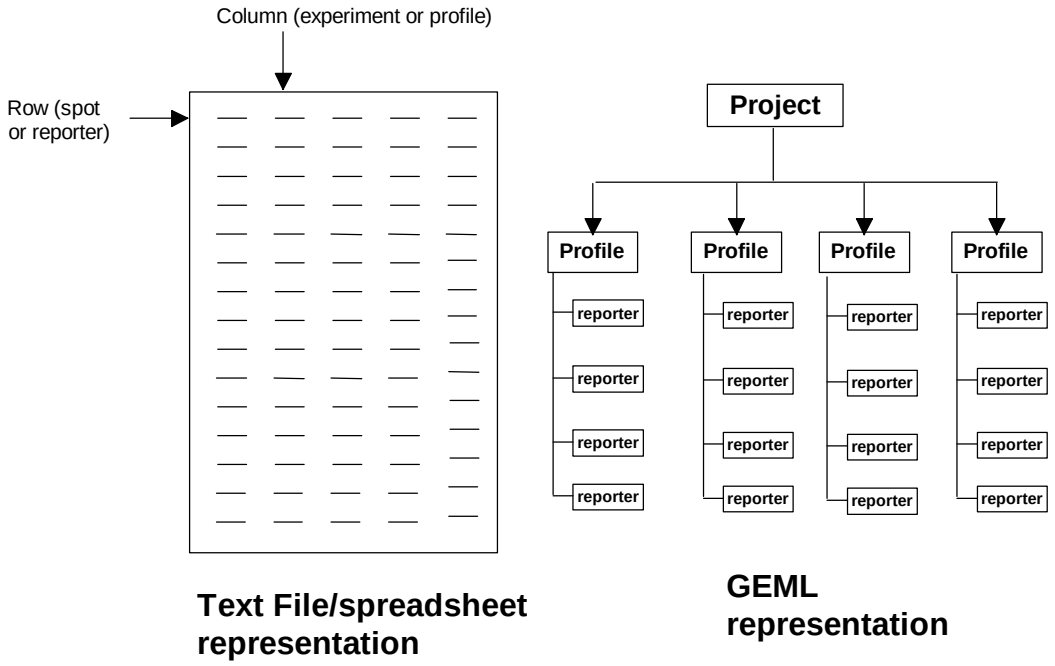

Figure 2: Standard Model for Biochip Experimental Data

The final objective for the biologist in conducting biochip experiments is to associate over-expression or under-expression of genes and proteins with biological processes and diseases. Thus, the Bioinformatics solution must provide information about gene and protein expression and be able to annotate genes and proteins with analysis results as well as related biological phenomena.

Each gene or protein should be annotated with biological information such as **chromosomal location, nucleotide sequence, promoter, protein sequence, protein domain and protein family, protein structure, protein function, biological pathway, etc.** Much of this

information is available in different databases, which are freely accessible over the Internet. However, using this information in the context of biochip experiments is a non-obvious process as:

1. The standard data model and the gene expression markup language is geared towards storing the primary numerical data from biochip experiments. There is no data model currently available for providing the user with a unified view of the expression data and related biological processes and diseases associated with genes and proteins
2. There is no single data source. The information required for annotation is available from heterogeneous and distributed data sources. The data sources are heterogeneous in their format, their access protocols and data base management systems, and also in their update schedules.

The solution of the common data representation proposed in this paper is unique and innovative in that:

1. It will allow a single user-friendly access to complete and necessary information in biochip experiments.
2. The unified data model is not specific to data sources but instead based on the idea of "annotation segments". The common representation does not merely concatenate information from a multitude of disparate databases. It cross-references and collates information from multiple databases into biologically relevant data items so that data in the common representation is irredundant.
3. The unified data model arranges the annotation information in a compact and sophisticated hierarchy that reflects the relations among the biological data items and is most suited to return answers to complex queries.

Definition: An annotation segment of the data model is defined as a data item or a collection of data items that relate to a single biological feature or to the results of a single analysis.

Biological feature segment example: There are four major repositories of nucleotide sequences - DNA Data Bank of Japan, EMBL Nucleotide Sequence Database, Genbank and the Genome Sequence Database at NCGR. The common data representation of this paper however provides a single "biological segment" called DNAseq for storing nucleotide sequence data. Thus, the representation for nucleotide sequence data is independent of which of the four source databases was used.

Table 1 below shows some further examples of annotation segments and the corresponding possible sources of data for each segment. If the database is freely accessible on the Internet, the corresponding URL is also provided. The table is not a complete list of all possible biological segments. Also, for each biological segment, the list of source databases or applications is not an exhaustive list.

Table 1: Annotation Segments for Biochip Experimental Data

Data Model: Biological Segment	**Source Databases**	**URL (All links valid until May 10, 2001)**
Expression Analysis	1. Partitioning Methods: K-means, Self Organizing Maps	K-means: Hartigan et al (1979) Self Organizing Maps: Tamayo et al (1999)
	2. Hierarchical Clustering	Murtagh (1983)

	3. Two-way Clustering	Eisen et al (1998), Getz et al (2000)
	4. Overlapped Clustering: Plaid Models, Fuzzy Clustering, Bayesian Methods	Plaid models: Lazzeroni and Owen (2000) Bayesian methods: Efron et al (2001)
	5. Classification: Permutation methods using t-statistics, Nonparametric rank statistics, and mutual information measures	Dudoit et al (2000) Pavlidis et al (2001)
Location	1. Yeast	ftp://genome-ftp.stanford.edu/pub/yeast/tables/ORF_Locations/ORF_table.txt
	2. Human (Cytogenetic Map) Human (Physical Map)	http://gdbwww.gdb.org/gdbreports/GeneByAlpha.tab http://www.ncbi.nlm.nih.gov/
	3. Mouse (Cytogenetic and Physical Map)	http://www.informatics.jax.org
	4. E. Coli (Physical Map)	ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/Ecoli/
DNA: Sequence	1. DNA Data Bank of Japan	http://www.ddbj.njg.ac.jp
	2. EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl.html
	3. Genbank	http://www.ncbi.nlm.nih.gov
	4. Genome Sequence Database	http://www.ncgr.org/research/sequence
DNA: Promoter	1. TRANSFAC	http://transfac.gbf.de/TRANSFAC/
	2. Promoter	http://www.epd.isb-sib.ch/
Protein: Sequence	1. Swiss-Prot/TrEMBL	http://www.ebi.ac.uk/swissprot/
	2. Genbank	http://www.ncbi.nlm.nih.gov
Protein Features, Motifs, Families, Domains	1. InterPRO	http://www.ebi.ac.uk/interpro/
	2. MetaFam	http://metafam.ahc.umn.edu/
Protein: 3D structure	1. Protein Data Bank (PDB)	http://www.rcsb.org/pdb/
	2. SCOP	http://scop.berkeley.edu/
Classification	1. Genetic Ontology Classification	http://www.geneontology.org
	2. COG	http://www.ncbi.nlm.nih.gov/COG/
Keywords	1. MeSH	http://www.nlm.nih.gov/mesh/meshhome.html
	2. Swiss-Prot	http://www.ebi.ac.uk/swissprot/
Function: Pathway	1. Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.ad.jp/kegg/
	2. WIT2	http://wit.mcs.anl.gov/WIT2/
Function: Gene-disease relation	1. OMIM	http://www.ncbi.nlm.nih.gov/OMIM
	2. Swiss-Prot	http://www.ebi.ac.uk/swissprot/
Citation: Bibliographic References	1. Pubmed	http://www.ncbi.nlm.nih.gov/Pubmed
	2. National Library of Medicine	http://www.nlm.nih.gov/nlmhome.html
	3. Swiss-Prot	http://www.ebi.ac.uk/swissprot/

The common data representation for the annotation information is organized into a data hierarchy. Some of the key elements of the hierarchy are depicted in Figure 3 which clearly shows in contrast its difference to the data hierarchy depicted in Figure 2. The common data representation and the data hierarchy are completely described in Table 2. The first column of Table 2 shows the different globally defined elements or data items; the second column provides a definition of the global element; the third column describes the hierarchical structure by indicating the children of each global element; the fourth column defines attributes: local data items that are used in the context for each global element. Attributes with the same name can have different meanings when associated with different global elements. The contextual meaning of each attribute is described in the details section of this paper.

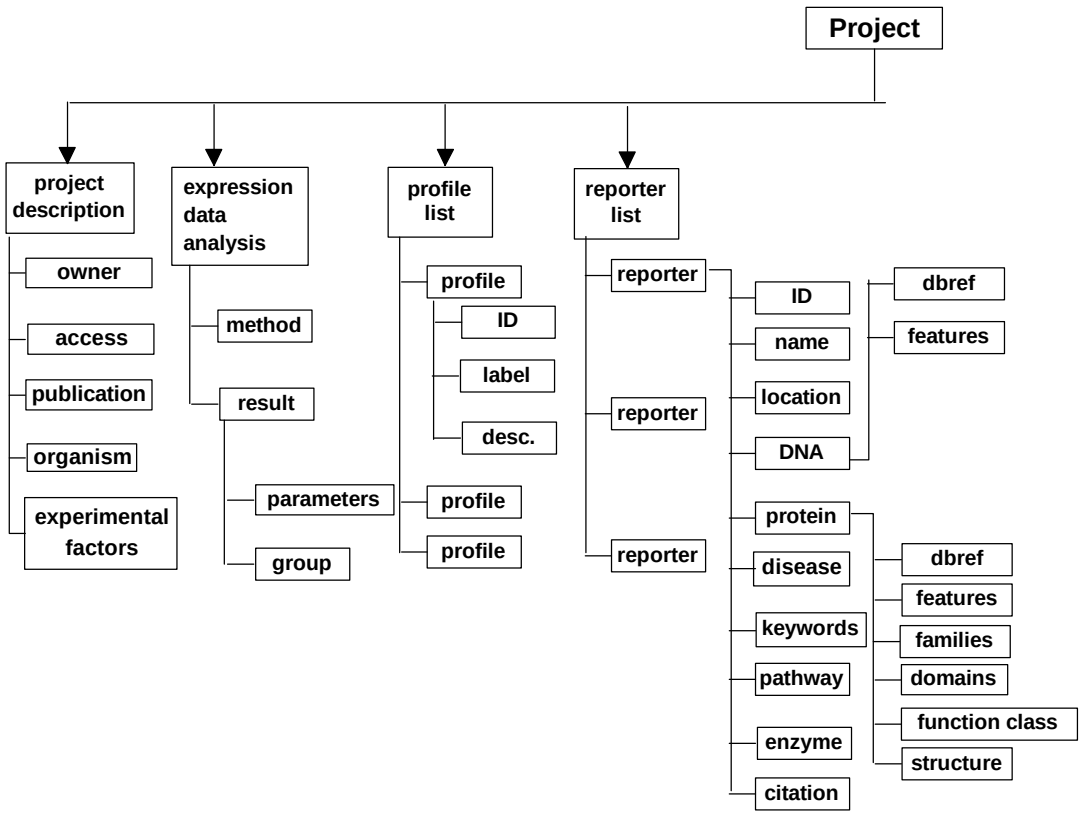

Figure 3: Some key elements of the common representation of annotation information for biochip experiments

Table 2: Complete Listing of Elements in Data Hierarchy of the Common Representation of Annotation Information

Elements	Definition	Children Elements	Attributes
project	Project is a set/series of biochip experiments performed for a single organism.	project_description(1), expression_analysis(*), profile_list(1), reporter_list(1)	id, name, version, date_created, date_released, date_updated
project_description	Parameters and descriptions relating to the entire project	owner(*), access(+), publication(*), organism(1), experimental_factors(?).	experiment_type

owner	Owner(s) of the Project	contact(+)	type
contact	Contact information, with a single entry per organization	individual_name(+), organization(1)	mailing_address, email, phone, fax
individual_name	Name of the individual(s) concerned		last_name, first_name, middle_name
organization	Institution information		name, laboratory, division
access	owner(s) of project may use this to control access to the data to groups or individual users	contact(*)	permission_type
publication	Stores information on publications/citations	dbref(?), article(?)	url
dbref	Stores information on publications/citations		db_name, db_id, comment
article	Information on the article published	author(+), journal(1)	title, abstract,, url
author	author(s) names and affiliations, with one entry per author.	individual_name(1), organization(*)	
journal	Journal information for publication		journal_name, year, volume, issue, pages
organism	Organism for which biochip experiment is conducted	genome(?)	common_name, biological_name, taxonomy
genome	Selected information about the genome of the organism	Chromosome(*), other(*)	
chromosome	Information on chromosome - in case of eukaryote organisms		number, ideogram_image, cM_length, bp_length
other	A catch all for terms not defined.		attribute_name, attribute_value
experimental_factors	A catch all for describing experiment specific paramters that may affect experimental outcomes	factor(*)	
factor	Description of a single experimental factor		name, value, comment, other
expression_analysis	To store the results of analysis of the microarray expression data	method(1), result(+)	id
method	Description of the analysis method	publication(*)	id, type, name, url
result	Designed to store the results from a large class of analysis algorithms	parameter_list(1), group(+)	
parameter_list	List for paramters used to obtain the result.	metric(1), parameter_other(*)	data_transform
metric	Distance/Similarity Metric used for expression data		type, name, formula
parameter_other	A catch all for method-specific		parameter_name, parameter_value

	parameters used to obtain the results		
group	"Groups" as obtained by the analysis.	children_id_list(?), reporter_id_list(?), profile_id_list(?), pair_id_list(?), annotation(*),	Id, group_score, group_significance_level
children_id_list	List of Ids of the groups that are children of the current group - used in hierarchical methods		children_id, membership
reporter_id_list	List of Ids of the reporters in the current group		reporter_id, membership
profile_id_list	List of Ids of the profiles in the current group		profile_id, membership
pair_id_list	List of Ids of the reporter-profile pairs in the current group		reporter_id, profile_id, membership
annotation	Annotation of group with relevant and significant biological information, obtained from different databases and applications.		Source, name, value, comment, other
profile_list	This is to store the list of profiles, where each "profile" is a single biochip hybridization.	profile(+)	
profile	A single biochip hybridization experiment	profile_description(*)	id, labels
profile_description	A catch all for profile description terms not already captured by the labels and the ID		name, value, comment, other
reporter_list	List of reporters in the set of biochip experiments.	reporter(+)	
reporter	A biological sequence used to measure a gene or protein's expression level.	location(?), nucleotide(*), protein(*), disease(*), keyword_list(?), pathway_list(?), enzyme_list(?), publication(*)	id, name, aliases
location	Cellular location for the reporter	dbref(?), position(?)	cellular_component, chromosome_number
position	Primarily used to indicate chromosomal position.		type, units, value, start, end
nucleotide	Stores nucleotide(DNA) information about the reporter	dbref(?), feature(*)	Length, sequence_type, sequence
feature	This is a catch all for storing annotation features		name, value, start, end, commnet, other
protein	Protein(s) associated	dbref(?), feature(*),	length, sequence

	with the reporter	protein_family(*), protein_domain, protein_classification, protein_structure	
protein_family	Categorization of closely related proteins into classes	dbref(?), protein_pattern(*)	family_name, superfamily
protein_pattern	A catchall to describe different patterns in proteins	dbref(?)	name, pattern, description. keywords, other
protein_domain	A discrete portion of a protein with its own function	dbref(?), protein_pattern(*)	name
protein_structure	The 2D/3D structure of a protein	dbref(?), protein_pattern(*)	type, description, image_url
protein_classification	Overall classification of protein	GO(*), COG(?)	
GO	Genetic Ontology Classification of a protein		type, id, class_name, class_description
COG	Phylogenetic Classification of a protrin		id, class_name, class_description
disease	Diseases associated with the reporter	dbref(+)	name, icd_class
keyword_list	List of keywords associated with the reporter	keyword(+)	type, measure_type
keyword	Keyword in keyword_list	keyword_main(1), keyword_qualifier(*)	
keyword_main	Primary keyword		primary_name, measure
keyword_qualifier	Aliases/associated keywords		qualifier_name, qualifier_type, measure
pathway_list	List of pathways associated with the reporter	dbref(?), pathway(*)	
pathway	Pathway in pathway_list		Id, name, type, class, map_url, ec_number_list
enzyme_list	List of enzymes associated with the reporter	dbref(?), enzyme(*)	
enzyme	Enzyme in enzyme_list	reaction(*)	ec_number, enzyme_name
reaction	Reaction(s) associated with the enzyme	Substrate_list, product_list	id
substrate_list	List of start compounds of the reaction	compound(+)	
compound	Description of a chemical compound		id, cpd_names, cpd_fprmula
product_list	List of end compunds of the reaction	compound(+)	

***eXpressML*:** This is an XML (extensible markup language) embodiment of the common representation.

XML Document Type Definition (DTD): The XML DTD or tags for ***eXpressML*** is described below with annotations to describe the significance and relevance of each tag element and its attributes. The content model for an element describes the children elements (sub tags). Furthermore, we indicate the number of permissible copies of each child element: (1) indicates a single copy, (?) Indicates optional i.e. zero or one copy, (*) indicates zero or more copies, and (+) indicates one or more copies.

Project: Project is a set/series of biochip experiments performed for a single organism.
<project
id = , ID number for the project
name = , Name of the project.
version = , Version number, required to indicate substantial updates to the data
date_created = , Date on which the project was first created
date_released = , Date on which data is made available to users.
date_last_updated = Date of last update
>
Content model for project: project_description(1), expression_analysis(*), profile_list(1), reporter_list(1).

Project Description: Parameters and descriptions relating to the entire project
<project_description
experiment_type = A project can describe exactly one type of series of experiments: currently the different types allowed are the same as in NCBI's Gene Expression Omnibus - GEO (http://www.ncbi.nlm.nih.gov/geo): dose-response, time course, other ordered group, parallel sample groups, repeat sample groups, and equivalent probe groups, other groups.
>
Content model for project_description: owner(*), access(+), publication(*), organism(1), experimental_factors(?).

Owner: Owner(s) of the Project.
<owner
type = type of owner: currently can be individual, group, or organization.
>
Content model for owner: contact(+)

Contact: Contact information, with a single entry per organization.
<contact
mailing_address = , Mailing address for project related mailings
email = , e-mail address to which inquiries about project can be sent
phone = , Telephone number for inquiries about project
fax = Fax number for project related transmissions
>
Content model for contact: individual_name(+), organization(1)

Individual_name: Name of the individual(s) concerned.
<individual_name

last_name = , last name of the individual
first_name = , first name of the individual
middle_name = middle name of the individual
>

Organization: Information about the organization
<**organization**
name = , Name of the organization
laboratory = Name of laboratory(ies) within the organization that are involved
division = Division/Department/Section of the organization
>

Access: owner(s) of project may use this to control access to the data to groups or individual users
<**access**
permission_type = This can be set to public (all users allowed access), limited (access limited to certain groups or individuals), private (access restricted to owners).
>
Content model for access: contact(*)
<**contact**> tag is as defined previously

Publication: Stores information on publications/citations
<**publication**
url = Internet address for online supplement
>
Content model for publication: dbref(?), article(?)

dbref: Database reference
<**dbref**
db_name = , Database from which information was extracted
db_id = , Accession number for the entry in database db_name
comment = Comments from Curator
>

Article: Information on the article published
<**article**
title = , Title of the article
abstract = , Abstract of the article
url = URL for the article
>
Content model for article: author(+), journal(1).

Author: author(s) names and affiliations, with one entry per author.
<**author**>
Content model for author: individual_name(1), organization(*)
<**individual_name**> tag is as defined previously
<**institutuion**> tag is as defined previously

Journal: Journal information for publication
<**journal**
journal_name = , Name of the journal where published
year = , Year of Publication
volume = , Volume Number
issue = , Issue Number
pages = Page Numbers
>

Organism: Organism for which biochip experiment is conducted
<**organism**
common_name = , Common name of the organism
biological_name = , Biological name of the organism
taxonomy = taxonomy of the organism
>
Content model for organism: genome(?)

Genome: Selected information about the genome of the organism
<**genome**>
Content model for genome: chromosome(*), other(*)

Chromosome: Information on chromosome - in case of eukaryote organisms.
<**chromosome**
number = , Number of the chromosome
ideogram_image = , URL for ideogram image of chromosome
cM_length = , Length of the chromosome in centimorgan units.
bp_length = Length (number of basepairs) of chromosome.
>
Other: A catch all for terms not defined.
<**other**
attribute_name = , Name of the attribute
attribute_value = Value for the attribute
>

Experimental Factors: This is a catch all for other experimental descriptors that may impact the project. Examples include: protocol used in sample treatment; hybridization procedures and parameters like salt concentration, pH value, and temperature; scanner; type of microarray; treatment type; compound; labelling parameters, etc.
<**experimental_factors**>
Content model for experimental_factors: factor(+)

Factor: Description of a single experimental factor
<**factor**

name = , Name of the factor
value = , Value or level of the factor in the project
other = , A catch all for other descriptions of the factor
comment = Comments provided by owner
>

Expression Analysis: This is to store the results of analyzing raw expression data using clustering or classification or other analysis algorithms. A variety of microarray data analysis algorithms were reviewed and a universal set of tags were developed to allow representation of the results from a large number of the algorithms. A sample list of algorithms that were considered is given in Table 1. Some other representative algorithms considered are listed in http://linkage.rockefeller.edu/wli/microarray/ . General clustering and classification algorithms were also reviewed for this purpose (Jain et al, 1999)

<expression_analysis
id = ID number for analysis >
Content model for expression_analysis: method(1), result(+)

Method: Description of the analysis method
<method
id = , ID for the method used
type = , Type of method. Currently allows either clustering or classification
name = , Name of method. e.g. Single Linkage Hierarchical Clustering, Naive Bayesian Classification, etc.
url = URL for accessing application
>
Content model for method: publication(*).
<publication> Tag is as defined above

Result: Designed to store the results from a large class of clustering and classification and other algorithms.
<result>
Content model for result: parameter_list(1), group(+).

Parameter list: List for parameters used to obtain the result.
<parameter_list
data_transform = Transformation of raw expression data
>
Content model for parameters: metric(1), parameter_other(*).

Metric: Distance/Similarity Metric used for expression data.
<metric
type = , Type of metric. Currently allows distance or similarity types
name = , Name of the metric
formula = Formula for computing the metric
>

Parameter other: A catch all for method-specific parameters used to obtain the result.

<**parameter_other**

parameter_name = , Name of the parameter

parameter_value = Value of the parameter

>

Group: "Groups" as obtained by the analysis.

<**group**

id = , ID number for this group

group_score = , Score/Statistic/Aggregate measure for the group

group_significance_level = , Associated significance level for the group

>

Content model for group: children_id_list(?), reporter_id_list(?), profile_id_list(?), pair_id_list(?), annotation(*)

Children_id_list: List of Ids of the groups that are children of the current group - required for hierarchical clustering.

<**children_id_list**

children_id = , comma-separated group ID numbers in sequence

membership = comma-separated number indicating strength/degree/value of membership

>

Reporter_id_list: List of Ids of the reporters in the current group.

<**reporter_id_list**

reporter_id = , comma-separated reporter ID numbers in sequence

membership = comma-separated numbers indicating strength/degree/value of membership

>

Profile_id_list: List of Ids of the profiles in the current group.

<**profile_id_list**

profile_id = , comma-separated profile ID numbers in sequence

membership = comma-separated numbers indicating strength/degree/value of membership

>

Pair_id_list: List of repoter-profile pairs in the current group - required for coupled clustering methods.

<**pair_id_list**

reporter_id = , comma-separated reporter ID numbers in sequence

profile_id = , comma-separated profile ID numbers in sequence
membership = comma-separated Numbers indicating strength/degree/value of membership
>

Annotation: Annotation of group with relevant and significant biological information, obtained from different databases and applications.
<**annotation**
source = , Source of the annotation
name = , Name of the type of annotation
value = , The annotation
comment = , Curator's comments
other = catch all for others
>

Profile List:
<**profile_list**> This is to store the list of profiles, where each "profile" is a single biochip hybridization.
Content model for profile_list: profile(+)

Profile: A single biochip hybridization.
<**profile**
id = , ID number for a profile
labels = Label(s) assigned to the profile
>
Content model for profile: profile_description(*)
profile_description: A catch all for profile description terms not already captured by the label and the ID
<**profile_description**
name = , Name of the description
value = , The description
comment = , Curator's comments
other = catch all for others
>

Reporter list: List of reporters in the set of biochip experiments.
<**reporter_list**>
Content model for reporter_list: reporter(+)

Reporter: A biological sequence used to measure a gene's expression level.
<**reporter**
id = , ID number for a reporter
name = , Primary name for the reporter
aliases = List of aliases for the reporter
>

Content model for reporter: location(?), nucleotide(*), protein(*), disease(*), keyword_list(?), pathway(*), enzyme(*), publication(*)

Location: Cellular location for the reporter
<**location**
cellular_component = , Cellular component in which this reporter is located, such as nuclear chromosome, mitochondria, chloroplasts
chromosome_number = For reporters located in eukaryotic nuclear chromosomes.
>
Content model for location: dbref(?), position(?)

<**dbref**> tag is as defined above
Position: Primarily used to indicate chromosomal position.
<**position**
type = , Type of position - currently allows linkage, cytogenetic, physical, radiation hybridization, other
units = , Units in which distance is measured
value = **,** Value of the position in specified units
start = , Start position of the reporter
end = End position of the reporter
>

Nucleotide: Stores nucleotide (DNA) information about the reporter
<**nucleotide**
length = , Length of the sequence
sequence_type = , type of sequence: currently allowed to be upstream of reporter, downstream of reporter, reporter region.
sequence = DNA sequence associated with the reporter
>
Content model for nucleotide: dbref(?), feature(*)

<**dbref**> tag is as defined above.

Feature: This is a catch all for storing annotation features which are known or predicted to be present at particular positions in this DNA sequence
<**feature**
name = , Name of the feature. Allowed names are CDs, exon, intron, promoter, poly-A, repeat, mutation.
start = , Start position of this feature
end = , End position of this feature
value = , Other characteristics of this feature
comment = , Comment regarding the feature
other = Other property of the feature
>

Protein: Protein(s) associated with this reporter

<**protein**
length = , Length of the amino acid sequence of the protein
sequence = Amino acid sequence associated with the protein
>
Content model for protein: dbref(?), feature(*), protein_family(*), protein_domain(*), protein_structure(*), protein_classification(*)

<**dbref**> tag is as defined above
<**feature**> tag is as defined above

Protein Family: Categorization of closely related protein into classes (Bateman et al, 2000; Apweiler et al, 2000; Shoop et al, 2001)
<**protein_family**
family_name = , Name of the protein family
superfamily = , Superfamily of the protein family
>
Content model for protein_family: dbref(?), protein_pattern(*)

<**dbref**> tag is as defined above
Protein_pattern: A catchall to describe different patterns in proteins (Attwood, 2000).
<**protein_pattern**
name = , Name of the feature
pattern = , Pattern associated with the feature
description = , Description of the feature
keywords = , Keywords associated with the feature
other = A catch all for other associated terms
>
Content model for protein_pattern: dbref(?)
<**dbref**> tag is as defined above

Protein domain: A discrete portion of a protein with its own function. The combination of domains in a single protein determines its overall function (Corpet et al, 2000; Henikoff et al, 1999; Kriventseva et al, 2000).
<**protein_domain**
name = Name of the domain
>
Content model for protein_domain: dbref(?), protein_pattern(*)
<**dbref**> tag is as defined above
<**protein_pattern**> tag is as defined above

Protein structure: 2D/3D protein structure
<**protein_structure**
type = , Type of structure
description = , Description of the structure
image_url = URL for the image depicting structure

>
Content model for protein_structure: dbref(?), protein_pattern(*)
<dbref> tag is as defined above
<protein_pattern> tag is as defined above

Protein classification: Overall classification of protein.
<protein_classification>
Content model for protein_classification: GO(*), COG(?)

GO: Genetic Ontology Classification of a protein (Ashburner, et al, 2000)
<GO
type = , Type of Ontologic Classification - either cellular component, biological process, or molecular function.
id = , Genetic Ontology ID number
class_name = , Name of the class
class_description = Description of the class
>

COG: Cluster of Orthologous Groups - Phylogenetic classification of proteins (Tatusov et al, 2001)

<COG
id = , COG ID number
class_name = , Name of the class
class_description = Description of the class
>

Disease: Diseases associated with reporter
<disease
name = , Name of the disease
ICD_class = International Classification of disease
>
Content model for disease: dbref(+)
<dbref> tag is as defined above

Keywords: List of keywords associated with this reporter
<keyword_list
type = , Type/Source of keywords. Currently allows MeSH, SwissProt.
measure_type = Type of measure used. Currently allows frequency, gene-average**,** document-average
>
Content model for keyword_list: keyword(*)

Keyword: Keyword in keyword_list
<keyword>

Content model for keyword: keyword_main(1), keyword_qualifier(*)

Keyword_main: Primary keyword associated with the reporter
<keyword_main
main_name = , Name of the keyword
measure = Measure of the importance of the keyword
>
Keyword_qualifier: Associated keywords associated with keyword_main
<keyword_qualifier
qualifier_name = , Name of the keyword
qualifier_type = , Type of association with keyword_main
measure = Measure of the importance of the keyword
>

Pathway List: List of Pathways that are associated with the reporter (Kanehisa, 1999). This is organism specific pathway information.
<pathway_list>
Content model for pathway_list: dbref(?), pathway(*)

<dbref> tag is as defined above

Pathway: Pathway in pathway_list
<pathway
id = , ID number associated with the pathway
name = , Name of the pathway
type = , Type of pathway - currently allows metabolism or regulatory.
class = , Class of the pathway
map_url = URL for the pathway map
ec_number_list = List of enzymes in the pathway
>

Enzyme List: List of enzymes associated with this reporter
<enzyme_list>
Content model for pathway_list: dbref(?), enzyme(*)

<dbref> tag is as defined above

Enzyme: Enzyme in enzyme_list
<enzyme
ec_number = Enzyme Classification Number (IUBMB, 1992)
enzyme_name = Name(s) of the enzyme
>
Content model for enzyme: reaction(*)

Reaction: Reaction corresponding to enzyme number and the current pathway

<reaction

id = , ID number of reaction

>

Content model for reaction: substrate_list(?), product_list(?)

Substrate list: Substrates (start compounds) of the reaction

<substrate_list>

Content model for substrate_list: compound(+)

Compound: Chemical compound information

<compound

id = , ID number for the compound

cpd_names = Names of the compound

cpd_formula = Chemical formula for the compound

>

Product list: Products (end compounds) of the reaction

<product_list>

Content model for product_list: compound(+)

Compound: Chemical compound information

<compound> tag is as defined above

Publications: Publication associated with the reporter

<publication> tag is as defined above

References:

1. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM (2000). "InterPro--an integrated documentation resource for protein families, domains and functional sites." Bioinformatics, volume16(12), pp:1145-1150
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature Genetics, volume 25(1), pp:25-29
3. Attwood TK (2000). "The quest to deduce protein function from sequence: the role of pattern databases." International Journal of Biochemical Cell Biology, volume32(2), pp:139-155.

4. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL (2000). "The Pfam protein families database." Nucleic Acids Research, volume 28(1), pp:263-266
5. Chakaravarti A (1999). "Population genetics - making sense out of sequence." Nature Genetics Supplement, volume 21, pp:56-60
6. Corpet F, Servant F, Gouzy J, Kahn D (2000). "ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons." Nucleic Acids Research, volume 28(1), pp:267-269.
7. Debouck C and Goodfellow PN (1999). "DNA microarrays in drug discovery and development." Nature Genetics Supplement, volume 21, pp:48-50
8. Dudoit S, Fridlyand J, Speed T (2000). "Comparison of discrimination mehods for the classification of tumors using gene expression data." Preprint #576, Statistics Dept, UC Berkeley
9. Efron B, Tibshirani R, Storey JD, Tusher V (2001). "Empirical Bayes analysis of a microarray experiment." Preprint #11B/216, Dept of Statistics, Stanford University
10. Eisen MB, Spellman PT, Brown PO and Botstein D (1998). "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences USA, volume 95(25), pp:14863–14868
11. Getz G, Levine E, and Domany E (2000). "Coupled Two-Way Clustering of DNA microarry data." Proceedings of the National Academy of Sciences USA, volume 97(22), pp:12079-12084
12. Hartigan JA and Wong MA (1979). "A K-Means Clustering Algorithms." Applied Statistics, volume 28, pp:100-108
13. Henikoff S, Henikoff JG, Pietrokovski S (1999). "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations." Bioinformatics, volume 15(6), pp:471-479
14. IUBMB: International Union of Biochemistry and Molecular Biology (1992). Enzyme Nomenclature. Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Academic Press, Inc., London.
15. Jain AK, Murty MN and Flynn PJ (1999) "Data Clustering: A Review." ACM Computing Surveys, volume 31(3), pp:264-323
16. Kanehisa, M (1999). " KEGG: From genes to biochemical pathways." In "Bioinfomatics: Databases and Systems" (Letovsky, S., ed.), pp. 63-76, Kluwer Academic Publishers.
17. Kriventseva EV, Fleischmann W, Zdobnov EM, Apweiler R (2001). "CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins." Nucleic Acids Research, volume 29(1), pp:33-36
18. Lazzeroni L and Owen A (2000). "Plaid models for gene expression data." Preprint #211, Dept of Statistics, Stanford University
19. Murtagh F (1983). "A survey of recent advances in hierarchical clustering algorithms." Computer Journal, volume 26, pp:354-359
20. Pavlidis P, Weston J, Cai J and Grundy N (2001). "Gene functional classification from heterogeneous data." Proceedings of the Fifth International Conference on Computational Molecular Biology, April 21-24, 2001. pp:242-248.
21. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, and Weinstein JN (2000). "A Gene Expression Database for the Molecular Pharmacology of Cancer." Nature Genetics, volume 24, no 3, pp:236-244

22. Shoop E, Silverstein KA, Johnson JE, Retzel EF (2001). "MetaFam: a unified classification of protein families. II. Schema and query capabilities." Bioinformatics, volume 17(3), pp:262-271
23. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES and Golub, TR(1999). "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentation." Proceedings of the National Academy of Sciences USA, volume 96(6), pp:2907-2912
24. Tatusov, RL, Natale, DA, Garkavtsev, IV, Tatusova, TA, Shankavaram, UT, Rao, BS, Kiryutin, B, Galperin, MY, Fedorova, ND, Koonin, EV (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." Nucleic Acids Research, volume 29(1), pp:22-28