Bioscience Reports, Vol. 25, Nos. 1/2, February/April 2005 (© 2005)
DOI: 10.1007/s10540-005-2850-4

A Bioinformatics Perspective on Proteomics: Data Storage, Analysis, and Integration

Andreas Kremer,[1,3] Reinhard Schneider,[2] and Georg C. Terstappen[1]

The field of proteomics is advancing rapidly as a result of powerful new technologies and proteomics experiments yield a vast and increasing amount of information. Data regarding protein occurrence, abundance, identity, sequence, structure, properties, and interactions need to be stored. Currently, a common standard has not yet been established and open access to results is needed for further development of robust analysis algorithms. Databases for proteomics will evolve from pure storage into knowledge resources, providing a repository for information (meta-data) which is mainly not stored in simple flat files. This review will shed light on recent steps towards the generation of a common standard in proteomics data storage and integration, but is not meant to be a comprehensive overview of all available databases and tools in the proteomics community.

KEY WORDS: Proteomics; database; data model; meta-data; HUPO; PSI; integration.

ABBREVIATIONS: DIGE: Differential in Gel Electrophoresis; EBI: European Bioinformatics Institute; GEML: Gene Expression Markup Language; HUPO: Human Proteome Organisation; ICAT: Isotop-Coded Affinity Tag; ISB: Institute for Systems Biology; MAGE-OM/ML: Microarray Gene Expression-Object Model/Markup Language; MIAME: Minimum Information about a Microarray Experiment; PAGE: Polyacrylamide Gel Electrophoresis; PEDRo: Proteome Experimental Data Repository; PIR: Protein Information Resource; PSI: Proteomics Standard Initiative; SBEAMS: Systems Biology Experimental Analysis Systems; SIB: Swiss Institute of Bioinformatics; XML: eXtensible Markup Language.

PROTEOMICS, LESS GENES THAN EXPECTED...

The term proteome was first introduced by Marc Wilkins 10 years ago during a proteomics meeting in Siena (Italy) to describe the set of proteins encoded and expressed by the genome. Proteomics is the large scale identification and functional characterization of all expressed proteins in a given cell (in a given state), including all protein isoforms and modifications, the protein interaction networks, protein structure determination and high-order complexes of proteins (Tyers and Mann, 2003). Proteomics is advancing rapidly as the result of powerful new technologies and experiments yield a vast and increasing amount of data. Experimental infor-

[1]Sienabiotech S.p.A. Discovery Research, Via Fiorentina 1, 53100, Siena, Italy.
[2]EMBL, Meyerhofstr. 1, 69117, Heidelberg, Germany.
[3]To whom corresponcence should be addressed. E-mail: akremer@sienabiotech.com.

mation about protein occurrence, abundance, identity, sequence, structure, properties, and interactions need to be stored. There is currently no common standard or definition of the minimum set of information about a proteomics experiment established. As recently observed in the case of microarray experiments, for any further development and improvement of data analysis open access to results is essential. Fundamental issues such as biological variability, pre-analytic parameters and analytical reproducibility remain to be resolved. Consequently, the analysis of proteomics data is currently descriptive, often manual and relies mainly on expert opinions. In the future, databases used in proteomics will increasingly serve as knowledge resources and provide the repository for diverse information (meta-data), which will not only be limited to simple flat-files.

Since the official publication of the first draft of the human genome (Lander *et al*., 2001; Venter *et al.,* 2001) it is believed that there are between 30,000–40,000 genes. This is far less than the number originally predicted and means that genes must work in permutations and combinations, which invokes a network with exponentially increasing interactions. Regulatory networks in biology are accelerating networks, where operations are reliant in the integrated activity of any or all of the components' nodes (Mattick and Gagen, 2005). Additionally, simple binary classification often makes no sense in terms of the physiological network. In general, subtle differences in the affinity of an interaction will lead to the off and on switch in a network. This poses challenges for the current and future databases in life sciences.

Traditional Databases

Common objectives of the majority of databases are maximization of annotation, minimization of redundancy and integration or at least linkage to other databases. In other words, to provide a parts list which describes molecular and cellular features from the viewpoint of an evolutionary relationship. The representation of gene and genome sequence data is fairly well standardized and initially, the main efforts for protein sequence data have centred around the Swiss-Prot and TrEMBL databases as well as the Protein Information Resource (PIR). In 2002, the UniProt consortium was created by SIB, PIR and the EBI. These three collaborating bodies have pooled their resources to create "the world's most comprehensive catalogue of information on proteins" (Apweiler *et al*., 2004). The triad of databases that form the basis of UniProt are the Archive (UniParc), the Knowledgebase (UniProt) and the Non-redundant reference (UniRef) and the system is designed to give easy access to the available protein information. A second initiative, the protein interactions database, IntAct, is based on a standard system for the presentation and annotation of protein-protein interactions. It is a public repository for data from the various partners and the published literature and a portable version has been created to facilitate installations on different local systems and subsequent (easy) sharing of data (Hermjakob *et al*., 2004b).

Until recently, the main purpose of database similarity searches was to detect homologous sequences, regardless of the species or remoteness of the relationship. The goal was to infer similarity of function from similarity of sequences and/or to study the evolution of protein families and domains. In proteomics studies the goal is slightly different and therefore different strategies and tools are required. Statistical

significance is important, but not in the sense of the probability that two sequences are related by chance. The objective of a database search in the case of protein identification using mass spectrometry is to find an exact, or nearly exact, match between sub-sequences (peptide fragments) and those sequences encoded by the genome. Additionally it is often important to identify the presence or absence of a protein or even alterations of post-translational modifications (Boguski and McIntosh, 2003).

Information at the level of the proteome is critical for understanding a cellular phenotype and its implications for normal and disease conditions. Proteome alterations in disease processes occur in many different ways and some of these are not predictable from genomic analysis. It is clear that a better understanding of these alterations and their consequences within the biological context will have a substantial impact in medicine. The identification and development of biomarkers for diagnostics and early detection of disease is of high interest for the pharmaceutical industry. Moreover, the understanding of biological networks increasingly provides the rational basis for (preliminary) decisions on drug targets and target suitability (Hanash, 2003).

CURRENT EFFORTS IN PROTEOMICS DATABASES...

The number of proteomics databases available in the world-wide web is increasing (Crawford and Garrels, 2000; Wojcik and Schachter, 2000). Selected resources are listed in Table 1 and an up-to-date list of available resources can also be found in the recent Database Issue of Nucleic Acids Research (http://nar.oupjournals.org/content/vol33/suppl_1/index.dtl). Currently, the generation and the analysis of proteome data are becoming increasingly popular, and the field is moving towards high-throughput approaches.

Typical 2D PAGE databases store 2D gel images obtained from specific tissues or cells. Identified spots are highlighted and varying amounts of information about

Table 1. Protein Interaction Resources and some URLs of Interest

Database of interacting proteins (DIP)	http://dip.doe-mbi.ucla.edu
BIND	http://www.bind.ca
MIPS	http://mips.gsf.de
Protein–protein interaction database (IntAct)	http://www.ebi.ac.uk/intact
Human Protein Reference Database (HPRD)	http://www.hprd.org/
Human Protein Interaction Database (HPID)	http://www.hpid.org
HUPO	http://www.hupo.org/
HUPO PSI	http://psidev.sourceforge.net/ http://psidev.sf.net/
Index site:	http://www.expasy.ch/ch2d/2d-index.html
The Proteome Analysis DB:	http://www.ebi.ac.uk/proteome/
Swiss 2DPAGE	http://www.expasy.ch/ch2d/
2DWG Image Meta-database	http://www-lecb.ncifcrf.gov/2dwgDB/
PEDRo	http://sourceforge.net/projects/pedro http://pedro.man.ac.uk/
Open Proteomics Database	http://bioinformatics.icmb.utexas.edu/OPD/
Systems Biology Institute	http://www.systemsbiology.org/
SBEAMS	http://www.sbeams.org/

the identified proteins are provided. Most structured databases (Table 1) provide "clickable" map functionality as well: clicking on a spot leads the user to a protein "summary sheet", featuring experimental information, annotations and cross-references to classical protein databases. Gel image analysis may range from very basic to fairly complex, depending on how much automation is required. Currently, one major limitation for the development and scale-up of 2D PAGE based analysis is the gel comparison process. This bottleneck hinders data interpretation and analysis. Until the output images of different 2D PAGE experiments—yielding different gel shapes—can be reliable compared, tedious and time-intensive manual analysis remains necessary.

Today's principal protein databases emphasize molecular and cellular features and annotations and are not well suited to represent physiology. To retrieve groups of proteins based upon known pathways or functional classifications is currently not always possible in a reliable or satisfying manner. Reasons for this situation are the speed at which the proteomics field is evolving and the inherited dynamism which makes it difficult to define key data and the very complex meta-data needed for analysis. Furthermore, annotations about post-translational modifications are still rare and difficult to locate in a consistent way (Jung *et al*., 2001). There is also the challenge of distinguishing annotations merely based on predicted modifications (derived from protein motifs) from those based on direct experimental evidence. A more suitable database would enable to classify proteins from a functional, as well as an evolutionary viewpoint. Such a data repository would also contain values of protein concentrations measured for certain conditions (e.g. for cancer tissue at a certain stage) which could then be compared to concentrations found in non-disease tissue.

The aim of functional proteomics is to describe the function of a given protein based on the global pattern of its molecular interactions. Several techniques, including computational methods are available or start to emerge for the exploration of the "interactome" (defined as the pattern of interactions of a proteome (Xenarios and Eisenberg, 2001)). Proteins that interact or are part of the same complex are generally involved in the same cellular process. The inherent complexity of interaction data has led to the design of various data structures to store interaction information (Eilbeck *et al*., 1999; Bader and Hogue, 2000; Xenarios *et al*., 2000). Earlier interaction databases mainly provided a basic display of the alphabetical interaction list (with annotations or cross-references to other protein databases) and some basic query tools. More recent databases tend to structure the "interactome" model in order to offer real navigation tools. Especially large-scale two-hybrid screens have generated a wealth of information describing potential protein-protein interactions. When compiled with data from other sources and functional tests, a network of interactions among proteins and between proteins and other components of eukaryotic cells can be constructed. These networks can be used to deduce potential signalling pathways, interactive complexes and most importantly provide clues to the function of previously uncharacterized proteins (Tucker *et al*., 2001). The Pathway Resource List (PRL, http://www.cbio.mskcc.org/prl/index.php) available on the web, is a new database that contains information of 166 internet pathway resources including protein interaction databases. Due to the fact that many physical interactions are conserved between species, it should be possible to

infer information about human protein interactions (and hence protein function) using model organism protein-interaction datasets (Lehner and Fraser, 2004). There is also increasing interest in and need for applying proteomics to the identification of disease markers (see also Calvo *et al*., this issue). Not only the potential of mass spectrometry to yield comprehensive profiles of peptides and proteins in biological fluids without the need to first carry out protein separations has attracted interest. In principle, such an approach is highly suited for marker identification because of reduced sample requirements and high throughput (Hanash, 2003).

THE NEED FOR A COMMON STANDARD…

Recognizing the disparity in proteomics data storage formats, The Human Proteome Organization launched the Proteomics Standards Initiative in 2002 with the main aim ''to define community standards for data representation in proteomics to facilitate data comparison, exchange and verification''. Standards are currently developed for mass spectrometry and protein-protein interaction data, as well as a general proteomics format. (HUPO Proteomics Standards Initiative, PSI; (Hermjakob *et al*., 2004a)). In brief, the workflow consists of physical separation of samples by gel electrophoresis, size-exclusion and/or affinity chromatography, followed by mass spectrometric examinations and protein identification by bioinformatics analysis. Over the past few years the number and size of proteomic datasets composed of mass spectrometry-derived protein identifications reported in the literature have been growing dramatically. In part, the need for guidelines is also driven by the fact that a significant but undefined number of proteins being reported as ''identified'' in publications are likely to be ''false positives'' (Carr *et al*., 2004). It is almost always possible to match a MS/MS spectrum to a peptide in the database; the difficult part is validating that the match is correct. While the accepted standards for peptide identifications are MS/MS data, a significant portion of the ''proteomics community'' continues to employ peptide mass fingerprinting data for peptide identification.

A proposal for a standard representation of both, the methods used and the data generated in proteomics experiments was presented recently (Taylor *et al*., 2003; Garwood *et al*., 2004a). It is analogous to that of the minimum information about a microarray experiment (MIAME) guidelines for ''transcriptomics'' and the associated microarray gene expression (MAGE) object model including a extensible mark-up language (XML) implementation. This Proteome Experimental Data Repository (PEDRo) data model describes in an implementation-independent manner the data that are required to be captured from a proteomics experiment (both, results and meta-data). It does not only contain information on protein separation and identification, but includes also detailed descriptions of the experimental samples, the mass spectrometric analyses conducted, the conditions under which the measurements were taken, the equipment used for these measurements and the software used to perform protein identification. It can be regarded as a ''proof of concept'' model that covers most of the workflows which are currently followed in proteomics experiments. Its main intension is to share experimental results and not to provide a comprehensive query or analysis environment for proteomics data.

SOFTWARE...

If full use is to be made of such rich data models, they must be associated with comprehensive software tools for data capture, dissemination and analysis. It is crucial that software development is linked at an early stage through agreed documentation, XML-based definitions and controlled vocabularies that allow different tools to exchange primary data sets. First efforts are emerging regarding databases (Potter, 2001; Garwood *et al.*, 2004b) and systems biology software infrastructure (ISB, http://www.systemsbiology.org/). There are not many Proteome Data Repository implementations in mySQL or Oracle available. Oracle is used in many other business areas and also has advantages, especially regarding complex searches or analyses over datasets. For example, PEDRo has only limited support for relative protein abundance data (DIGE and stable isotope labelling strategies) partly due to the lack of generic constructs for representing relationships between different types of measurements (e.g., relative expression readings).

A repository with a more generic focus is the Systems Biology Experiment Analysis Management System (http://www.sbeams.org/). SBEAMS is an html-based relational database that allows integration of disparate data types such as from ICAT and cDNA experiments and was implemented to facilitate data management at the ISB. For the end user the result is an interactive html window accessible from any web browser. Since all data from each step of the experiment are 'warehoused' in a unified schema in the RDBMS, quality control and data analysis tasks are greatly simplified. It is planned that SBEAMS will be compliant with the emerging MAGE-OM/ML (Microarray Gene Expression-Object Model/Mark-up Language) specification (http://sourceforge.net/ projects/mged) that will combine all previous microarray standards such as MAML, GEML, and GeneXML.

A commercial package, ProteinScape™ was developed in collaboration with Bruker Daltonik and is used in the HUPO Human Brain pilot phase. ProteinScape™ is an integrated bioinformatics platform which handles all essential steps of a proteome study (Bluggel *et al.*, 2004).

The development of statistically sound methods for assignment of protein identity from incomplete mass spectral data is critical for automated deposition into databases. Currently, this is still mainly a manual and hence error-prone process. The knowledge of "lessons learned" from analysis of DNA microarray data, including clustering, compendium and pattern-matching approaches, should in principle be transferable to proteomics analysis. Still, populating a database *de novo* can take a long time, but creating subsequent data sets which share some aspects of the experimental set-up will then be significantly less time-consuming.

FUTURE DATABASES...

One general issue observed over the last two decades is the very often missing or incomplete information about the evidence for an annotation. Without a clear statement as to whether the provided information is based on direct experimental evidence or based on a theoretical model or prediction method, such an annotation can be counterproductive or even misleading. This became apparent when theoretical motif, structure and various functional prediction methods were applied and the

results added to database annotations. An example are annotations on post-translational modifications which are still quite rare; evidence for these annotations is difficult to locate in a consistent way (Jung *et al.*, 2001). There is a strong need in the near future for much more detailed meta-data describing the experiments itself and providing additional information on how certain values were generated. Examples of such meta-data would be the values of protein concentrations and other measurable attributes in an experiment compared with normal ranges of values in reference samples, or the description of the algorithms for the applied data normalization. The ability to identify proteins, which appear or are missing or changing their concentration is becoming crucial for the comparison of normal and disease conditions.

As mentioned before, today's sequence-centric databases emphasize molecular and cellular features and annotations and are not very well suited for addressing physiological aspects. Future databases with a focus on integrative biology must allow the exploration and retrieval of data from a physiological point of view. A user should be able to retrieve a group of proteins based on pathways or functional classifications, or identify proteins that appear, disappear or show changes in their abundance under certain conditions (e.g., disease state). To achieve this it will be necessary to define key data of an experiment and to provide the very complex meta-data needed for analysis.

TECHNOLOGY CHALLENGES FOR DATABASES...

The generation and analysis of proteome data are now widespread. As with other technologies applied and used in biology, it can be anticipated that emerging high-throughput approaches and techniques will continue to increase complexity. The lessons learned from genomics (parallelized, miniaturized and automated procedures) are already applied in proteomics and will provide the basis for the generation of highly reproducible results. For database providers the high volume generation technologies leading to ever increasing amounts of data to be stored call for standard ways to represent data, and an agreed minimum level of annotation. As seen in the area of expression analysis data this is urgently needed to facilitate the analysis, dissemination and exchange of proteomics data. Especially the exchange of data between different research groups can be seen as a current bottleneck in the expression analysis field. Exchange and re-analysis by different research groups is a necessary prerequisite for the development of more robust and better analysis tools. Especially in proteomics the challenge will be to distinguish between the biological variability and the analytical reproducibility.

NEED FOR INTEGRATION

One necessary step regarding the analysis of proteomics data beyond the current situation (were analysis still relies mainly on human expert interpretation) will be to access highly standardized heterogeneous external databases. Every dataset in research is of limited use unless it can be integrated dynamically with the diverse knowledge surrounding genomic sequences, proteins and disease. 'Intelligent data' approaches are now emerging based on object-oriented programming that are designed to address the complex proteomics challenges more efficiently (Hancock

et al., 2002). Figure 1 shows the complex relationship of questions which need to be addressed to various extends during the analysis of proteomics data.

Past and current issues regarding integrating life science databases in general were summarized in recent overviews (Eilbeck *et al*., 1999; Köhler, 2004; Searls, 2005). It is often necessary to integrate ''derived data'' such as conclusions and not only raw data. Also in this regard the aspect of data quality and its control are crucial since integration of data 'interpretations' made by experts carries the risk of inconsistencies and human bias. The current database concepts need to be improved in order to integrate complex information about cellular regulation, pathways, networks and cellular roles (Tucker *et al*., 2001). The combination of results from transcriptomic and proteomic experiments, combined with data generated from biomedical, genetic and metabonomic approaches, protein interaction studies, model organism biology and clinical analyses, will become a useful tool for the individual researcher to build and test hypotheses (Eilbeck *et al*., 1999; Basik *et al*., 2003; Hegde *et al*., 2003; Köhler, 2004; Searls, 2005). But practical and biological differences exist regarding profiling the transcriptome and proteome in parallel. While current databases are sufficient for analysis of a particular protein or small interaction networks, they are not as useful for the integration of complex information about cellular regulation, pathways, networks and cellular roles, and they lack coordination and the ability to exchange information between multiple data sources (Tucker *et al*., 2001). Thus, it is still necessary to significantly improve the way in which data from functional genomics approaches (e.g., microarray-based expression profiles, etc.) and proteomic results are integrated. The resulting comprehensive

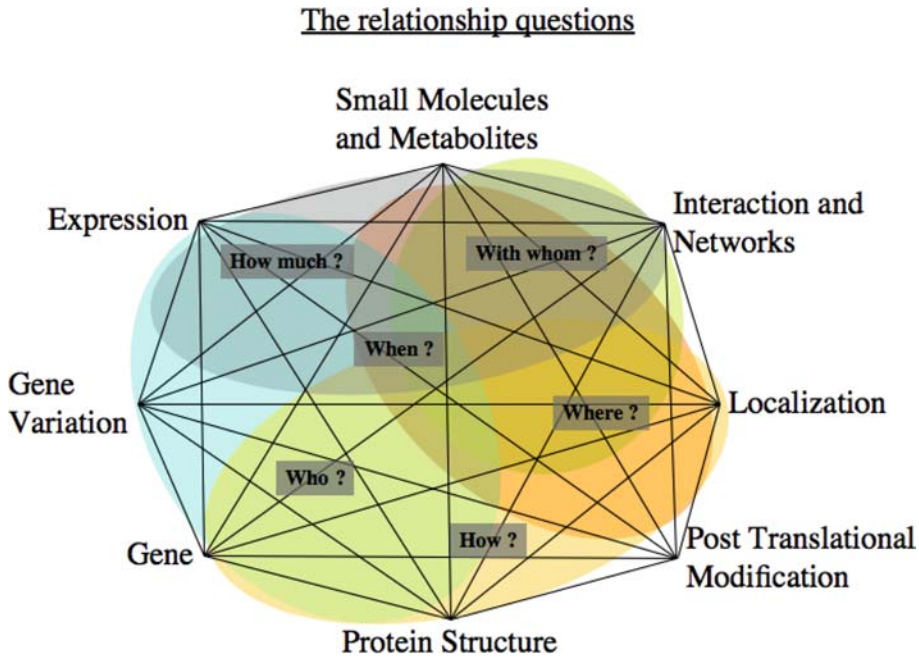

Fig. 1. The relationship questions in proteomics.

databases of gene function will serve as a powerful reference of protein properties and functions. The well-established relational database systems and flat-file retrieval systems (being commercial or open-source) are fundamental underlying technology stacks for bioinformatics. However, it is data format and exchange, rather than the database software itself, which needs attention. The wide use of XML as an exchange format and the increasing number of web-services will aid the process of integrating data from different sources.

A rich area for improvement is the field of terminology. Storing complex data with a wide range of concepts and models for the various aspects of a biological system is and will stay a challenge. Since the terminology in biology is in constant flow and still being developed, the situation of incompatible terminology and uncontrolled vocabulary is an issue. Despite the current efforts to establish a variety of ontologies, major problems in constructing uniform queries across many databases remain due to the lack of specifying the meaning and information content of a biological entity in an automatic fashion. In this respect, the "free-for-all' naming convention used by biologists is certainly an additional barrier (Petsko, 2002; Povey and Wain, 2002).

Information at the level of the proteome is essential for understanding the function of a cellular phenotype and its role in health and disease. The ongoing major collaborative efforts to collate and present publicly available proteomics data as well as efforts to standardize data formats will still need some time, but the improved understanding of biological networks will certainly result in greater knowledge of living systems and ultimately have a substantial impact in medicine.

REFERENCES

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32:** D115–D119 , Database issue.

Bader, G. D. and Hogue, C. W. (2000) BIND–a data specification for storing and describing biomolecular interaction molecular complexes and pathways. *Bioinformatics* **16**(5)465–477.

Basik, M., Mousses, S., and Trent, J. (2003) Integration of genomic technologies for accelerated cancer drug development. *Biotechniques* **35**(3) 580–2, 584, 586.

Bluggel, M., Bailey, S., Korting, G., Stephan, C., Reidegeld, K. A., Thiele, H., Apweiler, R., Hamacher, M., and Meyer, H. E. (2004) Towards data management of the HUPO Human Brain Proteome Project pilot phase. *Proteomics* **4**(8) 2361–2362.

Boguski, M. S. and McIntosh, M. W. (2003) Biomedical informatics for proteomics. *Nature* **422**(6928) 233–237.

Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol. Cell. Proteomics* **3**(6) 531–533.

Crawford, M. E., and Garrels, J. I. C. M. E. (2000) Databases and knowledge resources for proteomics research. *Proteomics: A Trends Guide*, pp. 17–21.

Eilbeck, K., Brass, A., Paton, N., and Hodgman, C. (1999) INTERACT: an object oriented protein-protein interaction database. *Proc. Int. Conf. Intell. Syst. Mol. Biol.***:** 87–94.

Garwood, K., McLaughlin, T., Garwood, C., Joens, S., Morrison, N., Taylor, C. F., Carroll, K., Evans, C., Whetton, A. D., Hart, S., Stead, D., Yin, Z., Brown, A. J., Hesketh, A., Chater, K., Hansson, L., Mewissen, M., Ghazal, P., Howard, J., Lilley, K. S., Gaskell, S. J., Brass, A., Hubbard, S. J., Oliver,

S. G., and Paton, N. W. (2004a) PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC. Genomics* **5**(1), 68.

Garwood, K. L., Taylor, C. F., Runte, K. J., Brass, A., Oliver, S. G., and Paton, N. W. (2004b) Pedro: a configurable data entry tool for XML. *Bioinformatics* **20**(15) 2463–2465.

Hanash, S. (2003) Disease proteomics *Nature* **422**(6928), 226–232.

Hancock, W. S., Wu, S. L., Stanley, R. R., and Gombocz, E. A. (2002) Publishing large proteome datasets: scientific policy meets emerging technologies. *Trends Biotechnol.* **20**(12), S39–S44.

Hegde, P. S., White, I. R., and Debouck, C. (2003) Interplay of transcriptomics and proteomics. *Curr. Opin. Biotechnol.* **14**(6)647–651.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004a) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22**(2), 177–183.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004b) IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32:**D452–D455, Database issue.

Jung, E., Veuthey, A. L., Gasteiger, E., and Bairoch, A (2001) Annotation of glycoproteins in the SWISS-PROT database. *Proteomics* **1**(2)262–268.

Köhler, J. (2004) Integration of Life Science Databases *Drug Discov. Today: BIOSILICO.* **2**(2), 61–69.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la, B. M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner,

L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., and Chen, Y. J. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860–921.

Lehner, B. and Fraser, A. G. (2004) A first-draft human protein-interaction map. *Genome Biol.* **5**(9)R63.

Mattick, J. S. and Gagen, M. J. (2005) Mathematics/computation. Accelerating networks. *Science* **307**(5711), 856–858.

Petsko, G. A. (2002) What's in a name? *Genome Biol.* **3**(4): COMMENT1005.

Potter, J. D. (2001) At the interfaces of epidemiology, genetics and genomics *Nat. Rev. Genet.* **2**(2), 142–147.

Povey, S.,and Wain, H. (2002), Smelling of roses? *Genome Biol.* **3**(6): INTERACTIONS1003.

Searls, D. B. (2005) Data integration: challenges for drug discovery *Nat. Rev. Drug Discov.* **4**(1), 45–58.

Taylor, C. F., Paton, N. W., Garwood, K. L., Kirby, P. D., Stead, D. A., Yin, Z., Deutsch, E. W., Selway, L., Walker, J., Riba-Garcia, I., Mohammed, S., Deery, M. J., Howard, J. A., Dunkley, T., Aebersold, R., Kell, D. B., Lilley, K. S., Roepstorff, P., Yates, J. R., Brass, A., Brown, A. J., Cash, P., Gaskell, S. J., Hubbard, S. J., and Oliver, S. G. (2003) A systematic approach to modelingcapturing, and disseminating proteomics experimental data. *Nat. Biotechnol.* **21**(3), 247–254.

Tucker, C. L., Gera, J. F., and Uetz, P. (2001) Towards an understanding of complex protein networks. *Trends Cell Biol.* **11**(3)102–106.

Tyers, M. and Mann, M. (2003) From genomics to proteomics. *Nature* **422**(6928), 193–197.

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., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di, F., , V, Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., and Nodell, M. (2001) The sequence of the human genome. *Science* **291**(5507), 1304–1351.

Wojcik, J. and Schachter, V. (2000) Proteomic databases and software on the web. *Brief. Bioinform.* **1**(3), 250–259.

Xenarios, I. and Eisenberg, D. (2001) Protein interaction databases. *Curr. Opin. Biotechnol* **12**(4)334–339.
Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**(1), 289–291.