The Use of Common Ontologies and Controlled Vocabularies to Enable Data Exchange and Deposition for Complex Proteomic Experiments

S. Orchard, L. Montecchi-Palazzi, H. Hermjakob, and R. Apweiler

Pacific Symposium on Biocomputing 10:186-196(2005)

THE USE OF COMMON ONTOLOGIES AND CONTROLLED VOCABULARIES TO ENABLE DATA EXCHANGE AND DEPOSITION FOR COMPLEX PROTEOMIC EXPERIMENTS

SANDRA ORCHARD
Sequence Database Group, EMBL-European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD UK

LUISA MONTECCHI-PALAZZI
Universita Tor Vergata,
Rome, Italy

HENNING HERMJAKOB
Sequence Database Group, EMBL-European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD UK

ROLF APWEILER
Sequence Database Group, EMBL-European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
CB10 1SD UK

Controlled vocabularies provide a roadmap through complex biological data. Proteomic data is increasing in volume and is currently poorly served by public repositories due to the large number of different formats in which the data is generated and stored. The Human Proteome Organization Proteome Standards Initiative is establishing standards for data transfer and deposition. These standards utilize ontologies and controlled vocabularies to describe experimental procedures and common processes such as sample preparation This paper will discuss the development of such ontologies by the user community and their current utilization in the fields of protein:proein interactions and mass spectrometry.

1. Introduction

Ontologies and controlled vocabularies are being established by many groups to provide roadmaps through the confused mass of data currently being generated from increasingly large-scale experimental biological experiments. The world of protein chemistry is no exception to this rule, with GO having lead the field by providing a framework in which individual molecules and complexes can be defined by their process, function and subcellular location [1]. The world's leading protein sequence database, UniProt [2], whilst incorporating and adding to the GO annotation of molecules described within UniProt-Swiss-Prot and UniProt-TrEMBL, also has its own defined keyword section that allows users to perform searches across the database using a standard nomenclature consistent to all entries. However, whilst the description of the function of these molecules is well served by established controlled vocabularies, the experimental techniques and procedures by which much of the functional information has been generated, has largely been ignored.

Proteomics is often described as the study of the protein translation products of the genome of a given organism but, in reality, this definition should be expanded to an understanding of the expression pattern and state of all proteins transcribed under a given set of conditions and the alteration of these parameters in response to a specific change to these conditions. The proteome of a cell encompasses the identity, subcellular location, post-translational modifications and protein:protein interactions made by the spectra of proteins expressed at any one moment in time and also how all these effect the function of both an individual protein and the cell as a whole. In order to map this, a multitude of experimental techniques have been developed. Proteins have first to be isolated and separated from a given biological sample, the latter usually either by 2-dimensional gel electrophoresis or by HPLC. The analytes are then ionized in the gas phase and the mass of the resulting peptide fragments measured by mass spectrometry. The resulting spectra are processed and specialized software used to match these fragments to known proteins. Such analyses will provide an expression map of the protein content of the cell under the defined experimental conditions – further techniques have been developed to provide further detail of the state of these proteins and their actual location within the cell. To fully understand the biological processes and pathways in which any one protein molecule may be involved, it is necessary to be aware of the interactions that molecule makes with other proteins, nucleic acids and small molecules within the cell. Experimental procedures by which these can be observed have been

established for many years but these too are becoming more high-throughput and the rate of data generation is rapidly increasing.

The context-sensitive nature of proteomic data necessitates the capture of a larger set of metadata than is normally required for sequencing, where knowledge of the organism of origin will suffice. Not only is information of sample source, handling, stimulation and eventual preparation for analysis required but also the detail of the analysis itself will need to be recorded. For example, to compare images of 2D-gels knowledge of their mass and charge ranges are required, and this information will need to be retrieved by users wishing to perform meaningful analysis of this experiment. Whilst the results and conclusion drawn from this data is frequently published in great detail, the underlying data is often only available as supplementary material, or is stored in author maintained databases or on websites. These databases and websites tend only to exist for the lifespan of the underlying project or grant, are often poorly maintained and the data within is difficult to access for downloading [3]. Published web addresses may lead nowhere [4]. Even when a stable database has been established, comparison between different datasets has proven difficult, in part due to the wide variety of terms and spellings used to describe a common experimental process. As an example, yeast two hybrid technology is a well known and widely used methodology for identifying protein interaction partners [5], a technique which has proved ideally suited to scaling up to increase throughput and data output. However, there are more than 10 different spellings for this term, e.g. Y2H, 2H, two-hybrid. While all of these are easily human understandable, non-standardized use of key terms makes systematic searches in large databases very difficult. Data interchange between databases sharing a common philosophy and even common data formats is being hampered by the lack of common terminology to describe identical processes. Even when a high level term can be used to describe a technique, for example mass spectrometry, data exchange and integration may require detail of instrumentation and data handling in order to give a complete picture of the conditions under which the data was generated – essential information for a full understanding of the importance of a particular dataset

2. The Human Proteome Organization Protein Standards Initiative

The Human Proteome Organisation (HUPO) was formed in 2001 with the aim of consolidating national and regional proteome organizations into a single worldwide body. The Proteome Standards Initiative was established by HUPO

with the remit of standardizing data formats within the field of proteomics to the end that public domain databases can be established where all such data can be deposited, exchanged between such databases or downloaded and utilized by laboratory workers [6]. The HUPO-PSI organized a series of meetings at which it was decided to develop a single data model that would describe and encompass central aspects of a proteomics experiment. This model would contain different sub-domains which will allow it to handle specific data types, for example 2-D electrophoresis gels or HPLC. Common processes would be described by a number of controlled vocabularies or ontologies. Where these processes are also relevant to micro-array data, for example in the area of sample preparation, this could be done in collaboration with the MGED consortium MGED (the micro-array gene expression data group), thus facilitating the comparison of proteomic with transcript data. Each sub-domain would then support a PSI-approved interchange format, which would permit the handling of data from many different sources. In the interests of making the task more manageable, the PSI agreed to concentrate their resources on two potential sub-domains, mass spectrometry and protein:protein interactions, whilst concurrently developing the encompassing proteomics data model, MIAPE [7].

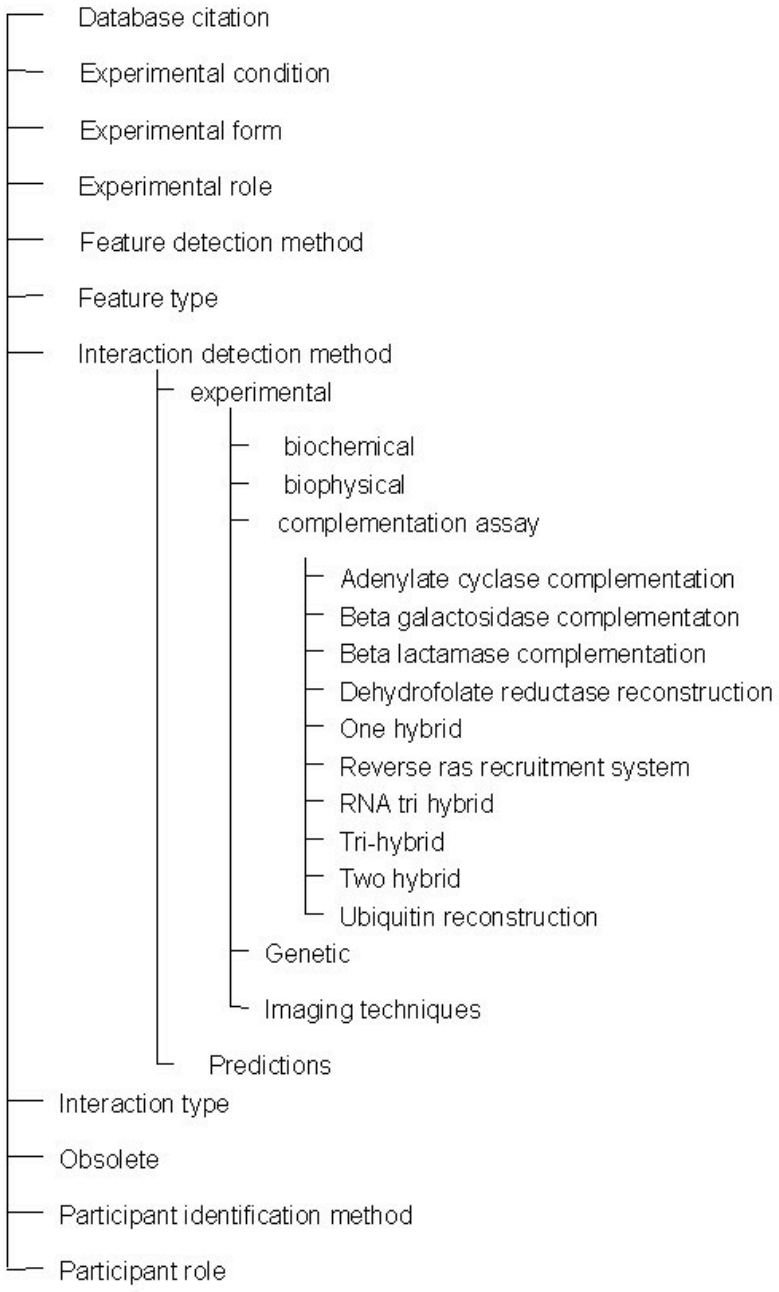
3. Molecular Interactions and the establishment of Common Vocabularies

A number of both commercial and academic molecular interaction databases already exist (IntAct [8], BIND [9], DIP [10], MINT [11], Hybrigencs [12], HPRD [13], MIPS [14]) which are wholly or partially in the public domain but, as discussed above, it was previously impossible to download or exchange data from any two of these databases in a common format. All these database providers, however, are committed to making their data more easily accessible and useful to the user community and actively supported the establishment of a common interchange format.

The HUPO-PSI MI format has been developed using a multi-level approach similar to that used by the Systems Biology Markup Language (SBML) [15]. Level 1, published early in 2004 [16], provided a basic format suitable for representing the majority of all currently available protein-protein interaction data. Level 2, released later in the year, allowed the description of protein interactions with nucleic acids or with small molecules. It allows the representation of both binary and n-ary interactions but does not contain detailed data on interaction mechanisms or full experimental descriptions. While a common data exchange format is a key requirement for an efficient exchange of

protein interaction data, it does not by itself guarantee data compatibility. It is essential to ensure standardized use of the data attributes through documentation and controlled vocabularies. The PSI-MI format contains detailed documentation within the XML schema itself, which is automatically extracted as an easily accessible web page and accompanied by detailed documentation on the HUPO-PSI MI context (http://psidev.sourceforge.net/mi/xml/doc/user/). To standardize the contents of data attributes, the PSI-MI format makes extensive use of controlled vocabularies or ontologies. External systems such as the Gene Ontology and the NCBI taxonomy, are referenced where possible. Detailed controlled vocabularies have been developed for the PSI-MI format for several key molecular interaction data attributes, such as the experimental methodology by which molecules are demonstrated to interact. Since this format already has a large user-base, it is intended to maintain the ontologies in GO format for the farseeable future, however it is recognized that, to become incorporated into the wider aspects of the General Proteomics Standards activities, it may become necessary to migrate to a more formal ontology language, such as OWL.

The PSI-MI format is designed for data exchange by many data providers. It is therefore important to ensure that both the data format (syntax) and the meaning of the data items (semantics) are consistent and well-defined. Without the standardization of data items as part of a community standard, data sets which are generated by the combination of data from different sources will quickly become difficult to search and to use. To address this problem, controlled vocabularies have been used in place of free text attributes, wherever possible. Several controlled vocabularies have been developed, including *interaction type*, *feature type*, *feature detection method*, *participant detection method*, and *interaction detection method* to describe specific aspects of both an interaction and the experimental methodology used to determine these (Fig. 1).



part of both the *interaction detection method* and *feature detection method* controlled vocabularies.

Fig. 1 Controlled vocabularies to describe experimental methodologies developed for use with the PSI-MI XML interchange format

technologies that can be used to infer that two or more proteins form a molecular aggregate. This vocabulary of more than 80 terms has a hierarchical structure based on a limited number of high level terms that group similar methods and reflect commonly used classifications and technical distinctions. As one method may be a specialization of more than one technology, a term may have more than one parent. For example, the “colocalization by fluorescent probes cloning” method (MI:0021) is both a fluorescence technology (MI:0051) and an imaging technique (MI:0428) The *participant detection method* controlled vocabulary lists more than twenty methods commonly used to establish the identity of the interacting partners, for example peptide mass fingerprinting (MI:0082).

The controlled vocabularies described here are not static; they will be maintained and updated by the HUPO PSI workgroup to reflect new experimental methodologies, or requirements from the community, in a manner similar to the maintenance of the Gene Ontology. This will ensure consensus on the inclusion of new terms by the user community, a high degree of flexibility to define terms when they are needed, and the avoidance of vague or ambiguous categories such as "other methods". In accordance with the GO model, an editorial team has been appointed [7] and requests for new terms dealt with via a SourceForge tracker system. Anyone interested in becoming more directly involved in the process is directed to the mailing list psidev-vocab@lists.sf.net.

Major interaction data providers are currently establishing the “International Molecular-Interaction Exchange (IMEx)” collaboration, in which they will regularly exchange user-submitted data, using the HUPO-PSI data exchange format. This will operate on similar principles to the EMBL/GenBank/DDBJ collaboration, thus providing a synchronized, stable, reliable resource for molecular interaction data. The success of this approach, with several of the above databases, for example IntAct, DIP. MINT and HPRD providing some or all of their data in PSI-MI format and others planning to follow suit in the near future, has provided encouragement for the progress of the related HUPO-PSI sponsored projects which are at an earlier stage of development due to the more complex nature of the data which they have to deal with.

4. Mass Spectrometry and the Establishment of Controlled Vocabularies

The HUPO-PSI mass spectrometry work group (PSI-MS) is working to develop a common data repository for the deposition of mass spectrometry data generated by proteomics groups and data standards accepted by both the user community and by instrumentation manufacturers. To this end, the group has produced the mzData format, a vendor-independent representation of mass spectra, providing a unified format for data archiving, exchange, and search engine input [7]. It has been jointly developed by academic users, commercial users and instrument vendors, among them Eli Lilly, EBI, Bruker Biosciences, Shimadzu, MDS Sciex, Agilent, and Thermo Electron. Controlled vocabularies will be used throughout mzData., in particular for the description of source detection methods, instrument parameters and analysis techniques. It has been proposed that the current ASTM mass spectrometry standard data dictionary be adopted, and updated, for use as a controlled vocabulary within this model, with eventual ownership of this dictionary potentially passing to the American Society for Mass Spectrometry (ASMS) so that it could be used to support both the HUPO-PSI and the ASTM's raw data standardization efforts. Users will also have the ability to develop their own vocabularies to allow for the specific needs of individual laboratories and maximum flexibility in experimental design. Work is currently ongoing to create and expand these vocabularies, in line with the final development of a beta version of mzData, a final release of which is planned for the HUPO world congress in Beijing, October 2004.

The mzData model will also act as the mass spectrometry component of the MIAPE data model and top level processes such as sample identification will be better dealt with through the development of this model than independently by the mass spectrometry group. The design of specifications for a spectral analysis format (mzAnalysis) is underway, and also that of a common syntax for the identification of proteins and peptides (mzProtID), which must also have the ability to describe post-translational modifications.

5. General Proteomics Standards and the Establishment of Controlled Vocabularies

As already stated, it is intended that all efforts within the remit of the HUPO-PSI will be coordinated and united within a framework provided by the establishment of standards for the representation of a full proteomics experiment, the Global Proteomics Standards (GPS). Based on the PEDRo

schema [18], this work group is tasked with developing the "Minimum Information About a Proteomics Experiment (MIAPE)" document analogous to the MIAME requirements for a micro-array experiment [19], and both an object model (PSI-OM) and XML format (PSI-ML) to fully represent a proteomics experiment. PSI-GPS will use the modules such as the more specific mzData format as components of a full experiment description, comprising sample preparation, analysis technologies, and results. To fully delineate these processes, controlled vocabularies are currently being written and appropriate terms will be contributed to the MGED Extended ontology under the "PSI" namespace. The MGED ontology is being written to support the micro-array object model, MAGE. The extended version adds further associations and classes to the core ontology which is intended to be stable and fully in synch with MAGE.

6. Summary

The design and use of controlled vocabularies to describe experimental data and enable the storage and exchange of proteomics data in a format that allows subsequent users a clear and comprehensive understanding of the conditions in which the experiments were performed, has only recently been tackled by workers in the field. The success of the HUPO-PSI MI format which uses a series of controlled vocabularies to describe molecular interactions, the features on a molecule responsible for such interactions and the experimental methods by which both interaction and features were determined suggests that ongoing work to describe detection of the proteome content of a sample by mass spectrometry will be equally successful. All these efforts should be seen in the wider framework of the GPS which is tackling more top level issues as sample description, acquisition and handling in conjunction with the MGED consortium, with the eventual aim of having a single combined ontology, with which to describe any experiment investigating the transcriptome and/or proteome content of a cell. This degree of cooperativity will ensure that these ontologies remain non-redundant and that the user community can access a single, common ontology suitable for describing complex experimental procedures.

It is always an issue, when launching a new standard, that this be seen as a fulfillment of a genuine need within the scientific community, rather than the bureaucratic imposition of unnecessary extra work without any perceived benefit. To this end, the HUPO-PSI proteomic standards are being written jointly by as

wide a cross section of the perceived end users as can practicably be achieved, with consultation at all stages of the process being an absolute requirement. The strides made within the protein:protein interaction community as a result of the publication of these standards and ontologies only a few months ago, lead us to hope that extending this process to cover the wider field of experimental proteomics will be equally productive and be of great benefit to an increased understanding of the proteome across all species and cell types.

References

1. M.A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, k. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G.M. Rubin, J.A. Blake, C. Bult, M. Dolan, H. Drabkin, J.T. Eppig, D.P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J.M. Cherry, K.R. Christie, M.C. Costanzo, S.S. Dwight, S. Engel, D.G. Fisk, J.E. Hirschman, E.L. Hong, R.S. Nash, A. Sethuraman, C.L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S.Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E.M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White; Nucleic Acids Res, **32** D258 (2004).
2. R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S. Yeh. Nucleic Acids Res. **32**:D115 (2004)
3. S. Orchard, C. Taylor, W. Zhu, R. K. Julian, Jr, H. Hermjakob, R. Apweiler Expert Review in Proteomics (in press)
4. J. Whitfield Nature. **428,** 592. (2004)
5. P. Legrain , L. Selig .FEBS Lett. **480,** 32 (2000).
6. S. Orchard, H. Hermjakob, R. Apweiler. Proteomics, **3**, 1374 (2003)
7. S. Orchard, C. F.. Taylor, H. Hermjakob, Weimin-Zhu, R. K. Julian, Jr, R. Apweiler Proteomics, (in press).
8. H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, R. Apweiler. Nucleic Acids Res., **32**, D452 (2004)
9. G.D. Bader, D. Betel, C.W.V. Hogue. Nucleic Acids. Res. **31**:248 (2003).
10. I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, D. Eisenberg. Nucleic Acids. Res. **30**:303 (2002).

11. A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, G. Cesareni.. FEBS Letts **513**: 135 (2002).
12. http://www.hybrigenics.fr
13. S. Peri, J.D. Navarro, T.Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T.K. Gandhi, K.N. Chandrika, N. Deshpande, S. Suresh, B.P. Rashmi, K. Shanker, N. Padma, V. Niranjan, H.C. Harsha, N. Talreja, B.M. Vrushabendra, M.A. Ramya, A.J. Yatish, M. Joy, H.N. Shivashankar, M.P.Kavitha, M. Menezes, D.R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C.K. Jonnalagadda, C.K. Prasad, C. Kumar-Sinha, K.S. Deshpande, A. Pandey. Nucleic Acids Res. **32** D497 (2004).
14. H.W. Mewes, C. Amid, R. Arnold. Nucleic Acids Res. **32** D41 (2004)
15. M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A. P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden , A.A. Cuellar, S. Dronov, E.D. Gilles, M. Ginkel, V. Gor, I.I. Goryanin, W.J. Hedley, T.C. Hodgman, J.H. Hofmeyr, P.J. Hunter, N.S. Juty, J.L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L.M. Loew, D. Lucio, P. Mendes, E. Minch, E.D. Mjolsness , Y. Nakayama, M.R. Nelson, P.F. Nielsen, T. Sakurada, J.C. Schaff, B.E. Shapiro, T.S. Shimizu, H.D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang;. Bioinformatics **19**:524-531 (2003)
16. H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik.,L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S.G. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, R. Apweiler. Nat Biotechnol**. 22** 177 (2004)
17. J.S. Garavelli Proteomics. **4**, 1527 (2004).
18. C.F. Taylor, N.W. Paton, K.L.Garwood, P.D. Kirby, D.A. Stead, Z. Yin, E.W. Deutsch, L. Selway, J. Walker, I. Riba-Garcia, S. Mohammed, M.J. Deery, J.A. Howard, T. Dunkley, R. Aebersold, D.B. Kell, K.S. Lilley, P. Roepstorff, J.R. Yates 3rd, A. Brass, A.J. Brown, P. Cash, S.J. Gaskell, S.J. Hubbard, S.G. Oliver SG. Nat Biotechnol. 2003, **21,** 24 (2003)
19. A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P.Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, M. Vingron M. Nat Genet., **29**, 365 (2001).