Letters

The importance of uniformity in reporting protein-function data

Rolf Apweiler[1], Athel Cornish-Bowden[2], Jan-Hendrik S. Hofmeyr[3], Carsten Kettner[4], Thomas S. Leyh[5], Dietmar Schomburg[6] and Keith Tipton[7]

[1]Head of Sequence Database Group, EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK CB10 1SD
[2]Institut Fédératif 'Biologie Structurale et Microbiologie', Bioénergétique et Ingéniérie des Protéines, Centre National de la Recherche Scientifique, 31 chemin Joseph-Aiguier, B.P. 71, 13402 Marseille Cedex 20, France
[3]Department of Biochemistry, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa
[4]Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Trakehner Str. 7–9, 60487 Frankfurt, Germany
[5]Department of Biochemistry, The Albert Einstein College of Medicine, Forchheimer Building, 1300 Morris Park Avenue, NY 10461, USA
[6]Universität zu Koeln, CUBIC – Institute of Biochemistry, Zuelpicher Strasse 47, 50674 Koeln, Germany
[7]Department of Biochemistry, Trinity College, Dublin 2, Ireland

The worldwide DNA-sequencing efforts annually deposit ~70 billion base-pairs of sequence information into the public databases – the equivalent of one *Escherichia coli* genome every 30 s – and the rate of deposition is doubling every 10–14 months. As biologists attempt to realize even a small fraction of the potential of this information by assembling accurate models of metabolic processes, they are thwarted by a surprising lack of uniformity in the acquisition and reporting of protein-function data – the data cannot be integrated into the models because the standards needed to link protein-function datasets to one another do not yet exist.

The public domain nucleic-acid-sequence and protein-structure databases are among the most comprehensive and influential databases in the biological sciences. Virtually every individual interested in a biological problem routinely uses these databases to enhance their perspective. Despite their enormous utility, the biological community has always recognized that sequence and structural information is not sufficient to describe biological processes – the dimension of protein function must be added. The interrelationships of sequence, structure and function are among the most powerful and synergistic in science. Databases representing two of the three aspects of this fundamental triumvirate have been established on a comprehensive, worldwide scale; it is time to begin to build the third.

The data-acquisition model used by the existing protein-function databases, such as KEGG (Kyoto encyclopedia of genes and genomes; http://www.genome.ad.jp/kegg), UniProt (universal protein resource; http://www.ebi.uniprot.org) and BRENDA (the comprehensive enzyme database; http://brenda.bc.uni-koeln.de), is retrospective – annotators retrieve data, often by hand, from the published literature and add it piece-wise into the database. A protein-function database capable of complementing the existing sequence and structural databases will require a prospective design that electronically captures data as it enters the literature. The submission of data for publication in a database-accessible format (which has worked well in other fields) would enable prospective data capture, but requires the creation of electronic data-submission forms that, importantly, will establish de facto standards for the reporting of protein-function data.

The automated acquisition and dissemination of scientific data have transformed the fields that have adopted such methods. It is now possible to imagine an environment in which all of the published behaviors of a given protein are integrated into the fabric of all relevant information and made available to theorists and experimentalists at the touch of a button – a truly breathtaking codification of scientific information.

The value of any database is dependent on the quality of the data that it contains. Whereas much structural data are independent of the conditions under which they are determined, this is not true for functional data. Enzyme, transport and interaction data involve parameters (e.g. K_d, K_m, K_i and rate constants) that are highly dependent on the conditions (e.g. temperature, pH, ionic strength and other system components) under which they are determined, in addition to the nature of the system being studied. Such values are of little use unless these conditions are clearly and fully stated. Furthermore, their use will be limited if the assay conditions are not standardized. For example, attempts to reconstruct the behavior of a metabolic pathway require that the parameters for all enzymes are obtained at the same pH value, which should correspond to the physiological pH of the compartment involved. Clearly, it would not be appropriate to formulate universal conditions for such parameter determinations because, for example, a temperature of 37°C might be appropriate for studying many mammalian systems, but it would not necessarily be suitable for studying the behavior of systems in poikilotherms (cold-blooded animals) or in extremophiles (microbes that inhabit extreme environments).

Corresponding author: Leyh, T.S. (leyh@aecom.yu.edu).
Available online 2 December 2004

The STRENDA (standards for reporting enzymological data) Commission has formed under the auspices of the Beilstein Institute (http://www.beilstein-institut.de) in response to the need for ensuring high and consistent quality for the reporting of functional data for enzymes and related proteins. Its mission is to define the requirements for the deposition of such data in a form that will then be available to any relevant database. Initially, it is the intention of the commission to produce checklists describing the data that should be made available when parameters are to be reported and recommendations for assay conditions for different systems. These lists will be made available for comment at the STRENDA website (http://www.strenda.org).

It is clear that the adoption of standardized data criteria and its deposition will require the support and input of the community, journals and international scientific-funding agencies. We hope that it will be recognized that, unless action is taken, the value of functional databases will be severely restricted.

0968-0004/$ - see front matter
doi:10.1016/j.tibs.2004.11.002