Published in final edited form as:
Nat Chem Biol. 2010 November ; 6(11): 785. doi:10.1038/nchembio.460.

A large-scale protein-function database

Rolf Apweiler[1], **Richard Armstrong**[2], **Amos Bairoch**[3], **Athel Cornish-Bowden**[4], **Peter J. Halling**[5], **Jan-Hendrik S. Hofmeyr**[6], **Carsten Kettner**[7], **Thomas S. Leyh**[8,*], **Johann Rohwer**[6], **Dietmar Schomburg**[9], **Christoph Steinbeck**[10], and **Keith Tipton**[11]

[1] Protein and Nucleotide Data Group, EMBL Outstation European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK [2] Department of Biochemistry and Chemistry, Vanderbilt University, Nashville, Tennessee, USA [3] University of Geneva, Geneva, Switzerland [4] Centre National de la Recherche Scientifique, Marseilles, France [5] Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, Scotland [6] Department of Biochemistry, Stellenbosch University, Stellenbosch, South Africa [7] Strenda Program, Beilstein Institute for the Advancement of Chemical Sciences, Frankfurt, Germany [8] Department of Microbiology and Immunology, The Albert Einstein College of Medicine, Bronx, New York, USA [9] Department of Bioinformatics and Biochemistry, Braunschweig Institute of Technology, Braunschweig, Germany [10] Department of Chemoinformatics and Metabolism, EMBL Outstation European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK [11] Department of Biochemistry, Trinity College, Dublin, Ireland

To the editor

In 2009, the *Journal of Biological Chemistry* published nearly 37,000 pages containing the data and analyses of biological entities. *Biochemistry* contributed 12,000 pages, the *Proceedings of the National Academy of Sciences* 22,568 and the *European Journal of Biochemistry* 7,446… Imagine if all of the protein-function data in those pages, and more, had been efficiently deposited to a database that was accessible, free of charge, worldwide. This is the primary objective of the Strenda Committee (Standards for the Reporting of Enzymological Data)[1–3].

The rate at which data is acquired frequently outstrips the capacity of the human mind to house it. Instead, we mine it. The ability to electronically cull the majority of mankind's knowledge of the functioning of a particular biomolecule at the push of a button would be an acutely effective, efficient research tool. Consider the benefits of crossing such information against single nucleotide polymorphism databases to identify the biochemical lesions associated with disease-linked mutations or associate the functional consequences of mutations with changes in the structures housed in the Protein Data Bank. Additionally, as systems biologists strive to integrate large swaths of metabolism, ready access to initial-rate equilibria and regulatory data will prove immensely useful. Perhaps the greatest value of such a database lies in the myriad ways in which it would integrate into the daily activities of individuals, worldwide. One cannot help but wonder what fraction of the protein-function literature is obscured or even lost to the researcher by imprecise search engines and retrieval strategies.

[*]Tom.Leyh@Einstein.yu.edu.

Competing interests statement
The authors declare no competing financial interests.

In October of 2003, a group of scientists gathered under the auspices of the Beilstein Institut for the Advancement of Chemical Sciences to address a problem of common interest: the large-scale collection of protein-function data. It was realized that such an endeavor would require developing community-based standards for the reporting of protein-function data and that an electronic form, acting as a portal for the deposition of data as it enters the literature, would provide a mechanism for the growth of a protein-function database to parallel the efforts of the scientific community.

Strenda has worked extensively with the scientific community to formulate recommendations to authors for the reporting of enzymological data (Box 1). These recommendations are the result of in-depth discussions that took place at each of five annual Experimental Standard Conditions of Enzyme Characterizations conferences—international meetings comprised of about 50 invitees from the academy, industry and editorial boards. It is hoped that the recommendations will prove an asset to authors and journals alike by clearly articulating the community standards; so far they have been adopted by 14 journals.

Box 1

Standards for the reporting of enzymological data

All reports of kinetic and binding data must include a description of the identity of the catalytic or binding entity (enzyme, protein, nucleic acid or other molecule). This information should include the origin or source of the molecule, its purity, composition and other characteristics, such as post-translational modifications, mutations and any modifications made to facilitate expression or purification. The assay methods and exact experimental conditions of the assay must be fully described if it is a new assay or provided as a reference to previously published work, with or without modifications. The temperature, pH and pressure (if other than atmospheric) of the assay must always be included, even if previously published. In instances where catalytic activity or binding cannot be detected, an estimate of the limit of detection based on the sensitivity and error analysis of the assay should be provided. Ambiguous terms such as ‘not detectable’ should be avoided. A description of the software used for data analysis should be included along with calculated errors for all parameters.

First-order and second-order rate constants should be reported in units of s^{-1} and $M^{-1}\times s^{-1}$, respectively. Equilibrium binding constants should normally be reported as dissociation constants with units of concentration (M, mM, μM, nM). The values k_{cat}, k_{cat}/K_M and K_m from steady-state enzyme kinetics should be reported in units of s^{-1}, $M^{-1} \times s^{-1}$ and concentration (mM, μM, nM), respectively. The steady-state specific activity of an enzyme should normally be reported as a k_{cat}. If there is considerable uncertainty in the molar concentration of the catalyst, the specific activity should be reported as a V_{max} (nmol, μmol) of product formed per amount of protein per unit time (for instance, $\mu mol \times mg^{-1} \times s^{-1}$).

The relationship of structure to function is among the most powerful in molecular science, yet an initiative to construct a protein-function database on the scale of the Protein Data Bank does not yet exist. The time for such an effort has come, and Strenda and the Beilstein Institute stand ready to assist in its implementation. Coupling the electronic submission of data contained in an article to its publication has been crucial to the development of the extant, large-scale databases. Toward this end, Strenda has developed an electronic data-entry form for the deposition of protein-function data that can be viewed at the Strenda website (http://www.beilstein-institut.de/en/projects/strenda/). We encourage readers to view the form and share their thoughts regarding its design and construction with the goal of developing it to the point at which it can be become part of our routine publication practices.

References

1. Apweiler R, et al. Trends Biochem Sci. 2005; 30:11–12. [PubMed: 15653320]
2. Armstrong RN. Biochemistry. 2008; 47:1–2. [PubMed: 18167079]
3. Taylor CF, et al. Nat Biotech. 2008; 26:889–896.