The Molecule Pages database

Joshua Li*, Yuhong Ning*, Warren Hedley*, Brian Saunders*, Yongsheng Chen*, Nicole Tindill†, Timo Hannay† & Shankar Subramaniam*‡

**San Diego Supercomputer Center, and ‡Departments of Bioengineering and Chemistry and Biochemistry, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA*
†Nature Publishing Group, The Macmillan Building, 4 Crinan Street, London N1 9XW, UK

The Alliance for Cellular Signaling (AfCS)–*Nature* Molecule Pages will be a comprehensive database of key facts about more than 3,000 proteins involved in cell signalling. Each entry will be created by invited experts and be peer-reviewed. Alongside the large-scale experiments being conducted by the AfCS scientists, the wealth of information contained in this database offers the potential of accelerating the pace of discovery in signal transduction research.

The understanding of cellular function and physiology requires insight into the complexity of the proteomic content of cells. The complete sequence of mammalian genomes provides a first sense of this content, but genome annotations do not capture many important aspects that contribute to the vast diversity of protein content and function. Splice variants and small variations in the genomic sequence translate into sequence polymorphisms in the proteome, which can alter protein function and contribute to pathophysiological conditions. More important, proteins display large repertoires of 'states', which contribute to the enormous diversity we observe in biochemical networks.

The state of a signalling molecule is characterized by covalent modifications of the native polypeptide, the substrates or ligands bound to the protein, its state of association with other proteins, and its location in the cell. A signalling molecule may be a receptor, a channel, an enzyme or another functionally defined class, and its state modulates its function. During signal transduction, a molecule may undergo a transition from one functional state to another.

Unlike the 'gene' parts-list of a cell, which can be obtained from high-throughput gene array measurements, the states of a protein can at present be ascertained only by painstaking biochemical and cellular experimentation. Our knowledge of the states of intracellular signalling proteins comes from detailed experiments and comparative analysis of cellular pathways across different species. But how can this vast amount of knowledge present in scientific literature be made easily accessible to the scientific community at large, rather than only to expert researchers whose research focus is the molecule of interest?

The Alliance for Cellular Signaling (AfCS)–*Nature* 'Molecule Pages' is a comprehensive database that will capture qualitative and quantitative information about a large number of signalling molecules and the interactions between them. It will be freely accessible from the AfCS–*Nature* Signaling Gateway web site (www.signaling-gateway.org/). The Molecule Pages will contain data from many public repositories in addition to information from published literature entered by expert authors. Authors will construct pages by entering information into web-based forms designed to standardize data input.

One of the principal barriers on constructing a database such as this lies in the complex and varied vocabulary used by biologists to define the attributes of a molecule. The database can be useful only if the information is described with a structured vocabulary along with well-defined relationships between the data 'objects' (for example, a protein sequence and a given modification). The building of this 'schema', or the database structure, is thus the first step towards the structured recording of available data about biochemical networks.

Database design considerations

The best way to structure the immense amount of knowledge associated with the states of a protein is to capture it in a relational database (one that records data in a series of tables in a way that avoids redundancy). This contains precisely defined data fields and precisely defined relationships between them represented by links between the tables.

The Molecule Pages database contains over 200 such relational tables. These define a myriad of parameters ranging from sequence to kinetic and thermodynamic parameters associated with molecular states and state transformations. The complete schema of the database (also known as the entity relationship diagram) is available on request from the AfCS bioinformatics group.

The Molecule Pages are freely accessible using any common web browser. The underlying system has a typical three-tier architecture. Behind the browser (or 'client') tier is a second ('application') tier composed of a collection of Java and Perl programs developed by the AfCS bioinformatics team and hosted on our servers; these provide the functionality of the system. The third ('data') tier, also hosted by the AfCS, uses an Oracle 9i relational database management system to store the raw data.

Description of the Molecule Pages database

The Molecule Pages database schema is divided into automated and author-entered data. The central part of the database is the molecule table, which defines each of over 3,000 AfCS proteins and forms an 'anchor' for both automated and author-entered data about each molecule. The primary element in the molecule table is the canonical mouse protein sequence, which is defined in an unambiguous manner from the corresponding GenBank sequences.

Automated data

The automated data component of each Molecule Page integrates information about that protein obtained from external database records. These include DNA and protein sequence information, structural information, sequence comparisons and related sequences, and basic biophysical and biochemical properties. These data come from SwissProt, GenBank, LocusLink, MGDB (Mouse Genome Database from Jackson Laboratories), Pfam, PIR, PRINTS, TrEMBL, TrEMBLnew, RefSeq, and the Interpro databases. These automated data are stored in relational tables, using the protein GI number (the GenInfo Identifier from GenBank) as the primary key. Links are provided to relevant National Center for Biotechnology Information (NCBI) database records. Each sequence can be imported into the Biology Workbench (http://workbench.sdsc.edu/) for further analysis. (The Biology Workbench is a web-based infrastructure that allows biologists to search and analyse many popular protein and nucleic-acid sequence databases.) The automated data are updated on a periodic basis.

Author-entered data

One of the main objectives of the Molecule Pages is to provide the community with information about the function of every protein

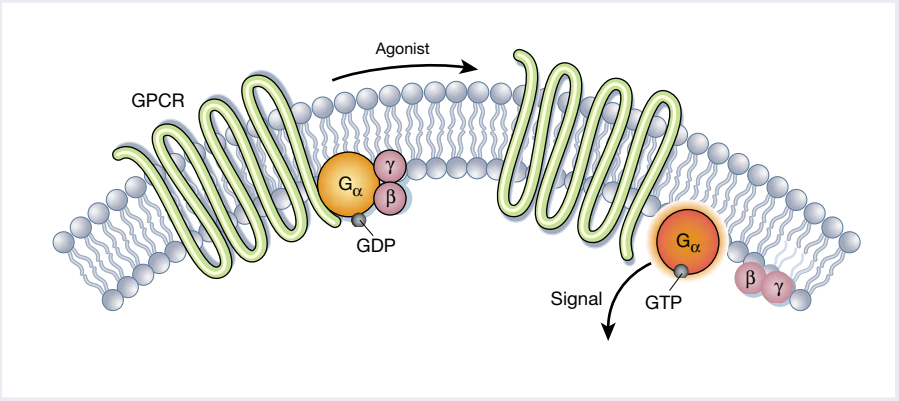

Figure 1 Example of molecule state changes. In unstimulated cells, the state of G_α is defined by its interaction with GDP, $G_{\beta\gamma}$ and a G-protein-coupled receptor (GPCR). Upon receptor stimulation by an agonist, its state is changed owing to its dissociation from the receptor and $G_{\beta\gamma}$, and the exchange of GTP for GDP, leading to G_α activation.

associated with cellular signalling networks, including qualitative and quantitative properties of each protein state. Such comprehensive, structured information about the states of signalling molecule proteins will aid the reconstruction — and hence the deeper understanding — of the signalling networks in which they participate. In addition, this information will help to provide the necessary data and parameters for quantitative modelling of signalling networks.

The primary role of an expert author is the entry and curation of these data in the Molecule Pages database. The author-entered data on a given state of the molecule are separated conceptually into two parts: the state of the molecule and the characterization of the function of that state. When describing the state, the author may invoke any number of possible modifications to the native polypeptide. These modifications include: the interaction of the protein with another protein, covalent modifications, binding to a substrate or ligand, and localization in a given subcellular compartment. In most cases a functional state of a protein will involve combinations of the above modifications.

The signalling protein G_α, for example, is in a defined functional state when bound to $G_{\beta\gamma}$, a G-protein-coupled receptor and GDP (Fig. 1). Agonist binding activates the receptor, which results in exchange of the substrate GDP by GTP, and leads to activation of the G protein. The author can define the bound state of G_α by defining the association with nucleotide GDP, $G_{\beta\gamma}$, and the receptor. The molecules with which G_α interacts are in turn defined by various functional states. For example, G_β is constitutively bound to G_γ. Even if no detailed information on $G_{\beta\gamma}$ has been entered previously into the database, the author of the G_α entry will be able to do so in a provisional form. At some time in the future, authors of G_β and G_γ entries will then be able to define the constitutive $G_{\beta\gamma}$ state more fully. In describing the function of G_α in this manner the author can define the activated state where the agonist is bound to the receptor followed by dissociation of G_α from the complexed state.

At each of these steps the author can enter any available quantitative data, such as binding constants or kinetic constants, along with the conditions under which the measurements were made. Detailed formats for entering such quantitative data have been defined in the Molecule Pages database. The above example shows how a complicated functional state of a molecule and its associated properties can be captured in tightly defined and well-structured database format. In this manner, the information can be put to use in the computer-assisted analysis and modelling of cellular signalling networks.

An important part of the Molecule Pages database is the information relating to functional alteration that arises from mutations of the components. For each AfCS molecule, we provide a format for defining all known sequence mutations, including natural variants, point mutations and deletions or insertions. In the definition of the state of a molecule, the author has the provision to indicate if the mutations alter the functional state and its properties. These data will provide valuable inputs into modelling the changes in cell signalling networks arising from mutations in component molecules and has potential implications in understanding the molecular basis of disease. The Molecule Pages database will also permit the author to provide a citation list pertinent to each entry and will link to the PubMed database maintained by the NCBI at the National Institutes of Health. Every Molecule Page will include an abstract providing a succinct overview of the characteristics and function of the protein in question. After each page has been created, peer reviewed and published, they will be updated annually to ensure that they remain current as more knowledge about a protein is gathered.

Relationship to AfCS experiments

Reconstruction of biochemical pathways is a complex task. In metabolic pathways, the task is somewhat simplified because of the essentially linear nature of the underlying processes, in which each step represents the enzymatic conversion of a substrate into a product. This is not the case in cellular signalling, where effects such as branching and cascades are commonplace.

The role of each protein in a signalling network is to communicate the signal from one node to the next, and to accomplish this, the protein has to be in a particular 'state'. We anticipate that the Molecule Pages will provide a catalogue of states for each significant signalling protein, such that one can begin to reconstruct signalling pathways with molecules in well-defined states functioning as nodes of a network. Interactions within and between functional states of molecules, as well as transitions between functional states, provide the building blocks for reconstruction of a signalling network. As described in the accompanying papers — see introductory article (pages 703–706) and articles on the signalling networks in B lymphocytes and cardiac myocytes (pages 708–710 and 712–714, respectively) — the experiments conducted by AfCS scientists will contribute to testing and validating such interactions and transitions in specific cells of interest. Together with the new large-scale experimental data sets being generated by AfCS laboratories, the highly structured review data provided by the Molecule Pages will, we hope, provide a new foundation for further accelerating the pace of discovery in cellular signalling, thus greatly enhancing our understanding of cellular processes in health and disease. □

doi:10.1038/nature01307

Correspondence and requests for materials should be addressed to S.S. (e-mail: shankar@ucsd.edu).