346–349 Nucleic Acids Research, 2001, Vol. 29, No. 1

Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems

Florence Horn*, Gerrit Vriend[1] and Fred E. Cohen

Department of Cellular and Molecular Pharmacology, UCSF, Box 0450, San Francisco, CA 94143-0450, USA and [1]CMBI, University of Nijmegen, Nijmegen, The Netherlands

Received September 7, 2000; Revised and Accepted October 19, 2000

ABSTRACT

The amount of genomic and proteomic data that is entered each day into databases and the experimental literature is outstripping the ability of experimental scientists to keep pace. While generic databases derived from automated curation efforts are useful, most biological scientists tend to focus on a class or family of molecules and their biological impact. Consequently, there is a need for molecular class-specific or other specialized databases. Such databases collect and organize data around a single topic or class of molecules. If curated well, such systems are extremely useful as they allow experimental scientists to obtain a large portion of the available data most relevant to their needs from a single source. We are involved in the development of two such databases with substantial pharmacological relevance. These are the GPCRDB and NucleaRDB information systems, which collect and disseminate data related to G protein-coupled receptors and intranuclear hormone receptors, respectively. The GPCRDB was a pilot project aimed at building a generic molecular class-specific database capable of dealing with highly heterogeneous data. A first version of the GPCRDB project has been completed and it is routinely used by thousands of scientists. The NucleaRDB was started recently as an application of the concept for the generalization of this technology. The GPCRDB is available via the WWW at http://www.gpcr.org/7tm/ and the NucleaRDB at http://www.receptors.org/NR/.

INTRODUCTION

Computational and experimental research on GPCRs and nuclear receptors benefits from the availability and easy accessibility of a substantial fraction of the information collected on these proteins. Information systems that codify this data should allow for the four basic data dissemination functions: browsing, retrieval, querying and inferring. Additionally, the information systems should provide tools for data gathering, data input, data validation, data annotation and manual curation. The data types, computational facilities and visualization tools of the GPCRDB were described extensively earlier (1,2) and will only be described briefly here. In an effort to build upon the GPCRDB software, we will describe our experience with data deposition, data handling and the design of molecular class-specific information systems (MCSISs) and highlight their application to the NucleaRDB project.

G protein-coupled receptors

G protein-coupled receptors (GPCRs) consist of a single protein chain that crosses the membrane seven times. These α-helical, transmembrane regions are presumably arranged in a fashion that is similar (3) to bovine rhodopsin (4). With the exception of this 2.8 Å bovine rhodopsin structure that became available recently, little atomic resolution structural information is available. GPCRs are of enormous importance for the pharmaceutical industry because ~50% of all existing medicines act on a GPCR (5). This has led many individuals within and outside the industry to build three-dimensional models of a variety of GPCRs. In all of these modeling studies, biochemical and pharmacological data, often obtained from the GPCRDB, was used to constrain the models and define the relative positions of residues.

Nuclear receptors

Nuclear receptors (NRs) are key transcription factors that regulate crucial gene networks important for cell growth, differentiation and homeostasis (6,7). Many of these receptors are potential targets for the therapy of diseases such as breast cancer, diabetes, inflammatory diseases or osteoporosis. NRs are grouped into a superfamily that includes receptors for steroid hormones, vitamin D, ecdysone, retinoic acid and thyroid hormone (see 6 for review).

NRs are modular proteins composed of six distinct regions (A–F) (6) that correspond to functional and structural domains. Not all the NRs contain all the six domains. Regions C and E display the highest degree of conservation. C is involved in DNA binding and E in ligand binding and dimerization. The C domain is the signature motif of the superfamily due to its high conservation. It is composed of two zinc fingers and the presence of this feature facilitates the identification of NRs (7). The superfamily has been subdivided into seven subfamilies (8,9) based on the alignments of the conserved domains. In contrast to GPCRs, a large amount of structural data is available for NRs.

Consequently, our understanding of the function of nuclear receptors is derived from the integration of a great deal of heterogeneous data. Sequence and structural data is available

*To whom correspondence should be addressed. Tel: +1 415 502 5650; Fax: +1 415 476 6515; Email: horn@cmpharm.ucsf.edu

in existing databases but there are only a few sources on the WWW for the other experimental data. For example, the Nuclear Receptor Resource (10) gathers different individual resources for glucocorticoid, androgen, thyroid hormone, Vitamin D and peroxisome-proliferator activated receptors.

CONTENT OF THE TWO MCSISs

Data types introduction

The information systems hold four types of experimental data: sequence, structure, mutation and ligand binding data. All four data types have their own specific problems. Sequences can be incorrect, truncated or can occur in several alternative translations. The interpretation of mutation studies and ligand binding studies depend strongly on experimental conditions such as cell type, stable versus transient transfection, density of receptors on the cell surface, second messenger assay and the kinds of ligands used.

There is no limit to the number of data types that can be derived computationally. It is therefore important to consider the questions that the systems should help answer when adding more computational data. Presently, the central computational data type is a multiple sequence alignment. Phylogenetic trees and correlated mutation analyses (CMA)(11) are derived from these multiple sequence alignments.

Visualization of data is important for users. In the GPCR field, two-dimensional representations by so-called snake-like diagrams are commonly used to visually combine a sequence with other types of information such as three-dimensional localization, mutation results, ligand binding or biochemical studies.

The two information systems are updated every month based on the availability of new data. The same software is used to maintain both systems. A detailed description of the methods used to update the GPCRDB has been already published (2). Table 1 summarizes the data content of the two systems.

Table 1. Contents of the GPCRDB and NucleaRDB information systems

Data types	GPCRDB (July 2000)	NucleaRDB (August 2000)
Receptor sequences	1796	630
including fragments	444	117
Families, subfamilies, …	222	109
Unclassified sequences	0	106
cDNA–protein alignments	2434	725
Multiple sequence alignments	213	96
3D models (nine depositors)	1290	*
Ligand binding data	~12 000	*

*Not yet available at the time of the preparation of this manuscript.

Primary data

Sequences. Sequences are automatically imported from the SWISS-PROT and TrEMBL databases (12). We are working on methods to harvest the large body of genome sequences (including human). Sequence fragments are deleterious for most computational purposes but they do hold useful information, and are thus stored separately. cDNA sequences are imported from the EMBL databank (13) via SRS (14).

Structural data. The systems provide the list of available PDB files and models built by homology as well as links to three web sites that display structural information: PDB Structure Explorer (15), the PDBREPORT database (16) (outputs of the WHAT_CHECK Structure Analysis) and PDBsum (17). The PDB identifiers are extracted from the database cross-references of the SWISS-PROT entries.

Mutation data. In the GPCRDB project, the mutation data is stored in an external database, tinyGRAP (18). The tinyGRAP database holds about 8300 mutants including single and multiple substitutions, chimeras, deletion and insertion mutations. No comprehensive mutation database exists for nuclear receptors with the exception of the androgen receptors (19).

Ligand binding data. Ligand binding data for GPCRs was obtained from P. Seeman (20). This very impressive collection of drug dissociation constants was manually extracted from the literature. Seeman collected data for neuroreceptors and transporters. About 12 000 dissociation constants are available for 28 GPCR families and subfamilies. Unfortunately, this data collection is not being maintained.

Computationally-derived data

Multiple sequence alignments. Multiple sequence alignments are performed with WHAT IF (21) as described by Oliveira *et al.* (11). These alignments are made for families, subfamilies, groups, etc. The alignments are presented in several formats.

cDNA–protein alignments. cDNA–protein alignments are generated using GeneWise, a component of the Wise2 package (http://www.sanger.ac.uk/Software/Wise2/), and checked and annotated for mismatches.

Phylogenetic trees. Phylogenetic trees are useful to visualize the relationships between sequences in a family. This information can help to answer several different kinds of questions, including those related to ligand design. While many algorithms have been developed to create phylogenetic trees, we have employed a neighbor-joining algorithm.

Correlated mutation analysis. Correlated mutation analysis is a computational method to identify pairs of sequence positions that remained conserved or mutated in tandem during evolution. The idea behind the search for such pairs of residues is that when a mutation occurs at a functionally important site, the protein either becomes non-functional or may acquire its original or a different function due to a compensatory mutation at another position. Residues detected by the CMA method are often involved in intermolecular interactions (between ligands and receptors or G-proteins and receptors; 22,23). A detailed explanation for this phenomenon is beyond the scope of this article.

Three-dimensional models. For each new family added to the system, the best template has to be selected manually and models are then built automatically for all sequences in the

family. The template selection and modeling will be automated in the future. For each model, one can either download the coordinates or view them using a WWW helper application like Rasmol (24) or WebMol (25).

Data visualization tools

Experimentalists in the GPCR field prefer to represent their data using two-dimensional snake-like diagrams. The Viseur program (26) can automatically generate snakes and hyperlink them to other types of information. Snakes are used in the GPCRDB to represent two kinds of data. One set of snakes is hyperlinked to the tinyGRAP mutant database. The second set is used to indicate the location of residues detected in the CMA analyses.

Inter-operability via database cross-reference tables

Database cross-reference tables are provided in the two information systems. A user-friendly view lists all the available pointers to local and remote information. Each pointer is hyperlinked to the corresponding data. This is done automatically by reading the SWISS-PROT entries and querying local and remote databases. Since August 2000, the SWISS-PROT entries for GPCRs and NRs point to these cross-reference tables.

The two MCSISs also contain pointers to other receptor-related WWW pages. This includes the addresses of GPCR specialists, pointers to external pages of different levels of relevance for GPCR and NR research (articles, group pages, GPCR related diseases, etc.) and other useful information resources.

DISSEMINATION FACILITIES

The MCSISs have been conceived to provide fast and easy access to all information related to the underlying molecular classes. For this purpose we have implemented (and are still implementing) the four basic information system tools: browsing, retrieval, query and inferring.

Browsing

The data organization is based on the pharmacological classification of receptors and access to the data is obtained via a hierarchical list of known families in agreement with this classification. For one specific family, one can access the individual sequences, the multiple alignments, the profiles used to perform the latter, two-dimensional visualization and a phylogenetic tree. Each type of data is displayed in a WWW page with hyperlinks to other data where appropriate.

Retrieval

Often a user wants to work on certain data independent of the information system environment. Therefore, most data can be retrieved in its native form using the 'save as' option of the WWW browsers or via anonymous FTP.

Query

In both MCSISs, a query system allows users to make simple queries via keyword or SWISS-PROT identifier and accession number.

In the GPCRDB, an advanced query system has been implemented in collaboration with a computer research institute (GMD, Darmstadt, Germany) that allow users to conduct simple and advanced queries, such as the search for a sequence pattern in a helix or a loop, by means of logical and regular expressions. The user can also refine the search by combining different queries. A fault tolerant query system automatically adjusts queries that lead to no hits. This adjustment can be linguistic (e.g. 'human' corresponds to 'homo sapiens') or relaxing (e.g. 'in helix III' corresponds to 'near helix III' or 'PPP' to 'PP').

In addition, a BLAST server at the European Bioinformatics Institute (EBI) allows the user to scan one sequence pattern against all the sequences stored in the GPCRDB.

Inferences

Besides structure analysis, correlated mutation analysis provides the most powerful tool available for the computational discovery of novel inter-residue relationships. CMA represents a powerful inference engine. With little computational effort, the potentially important residues are selected from among the (tens of) thousands of residues in each alignment. We make these residue positions available for browsing purposes by displaying them in two-dimensional plots with appropriate hyperlinks. Additional information about functionally relevant residues can be garnered from the use of the evolutionary trace algorithm (27).

DISCUSSION

Molecular class-specific information systems are among the few tools available to aid molecular scientists in managing the deluge of experimental data. The GPCRDB and the NucleaRDB are two MCSISs for G protein-coupled receptors and NRs, respectively. The interesting aspect of these two MCSISs is that they are both produced and maintained in a consistent way using software. This approach will allow us to produce MCSISs for a variety of other interesting classes of macromolecules. The only human intervention that is needed to create an MCSIS is the initial assignment of sequences in families, subfamilies, sub-subfamilies, etc. After that, the entire update procedure is automatic.

The major bottleneck in database maintenance is data entry or aggregation. Databases can only provide the user with information that has been entered and indexed into a computer file. Unfortunately data deposition is only obligatory for sequences and three-dimensional atomic coordinates. All other experimental data has to be manually extracted from the literature and entered into databases. Typically, this has been done by data managers and curators. In an anecdotal way, the GPCRDB project has shown that most experimental scientists are not yet ready to enter their own data through data input systems. Consequently, the next step in data handling must be the design of techniques for the automatic extraction of biologically relevant information from the literature.

The experimental data such as mutation data, ligand binding information, expression data, etc., will be automatically obtained by having computer software read electronically stored articles. Emphasis will be placed on canonical tables of mutagenesis experiments and their results, as well as structure activity tables that join chemical entities to biological impact with respect to enzymatic inhibition, receptor binding affinities or more integrated cellular readouts. The existing experimental data of the GPCRDB will provide a benchmark for the fidelity

of automated data extraction algorithms. Automated model construction by homology to proteins of known structure will be used to provide a structural context for other computational and experimental information.

The GPCRDB is now 4 years old and is used on average more than 2000 times per day. Initial interest for the NucleaRDB indicates that this site will become equally popular. The success of the GPCRDB and the interest in the recently configured NucleaRDB suggest that MCSISs are a useful solution for providing, disseminating and harvesting heterogeneous data.

Usage and availability

The MCSISs are accessible from http://www.gpcr.org/7tm/ and http://www.receptors.org/NR/, respectively. A European mirror is available for the NucleaRDB at http://www.gpcr.org/NR/. The underlying data files (alignments, models, etc) can be downloaded from anonymous FTP from ftp://www.gpcr.org.7tm/ and ftp://receptors.ucsf.edu/pub/NR/. Access to the MCSISs is free for academic and industrial scientists.

ACKNOWLEDGEMENTS

We thank K. Aberer, A. Bairoch, A. Bogan, M. Beukers, R. Bywater, F. Campagne and Ø. Edvardsen, E. L. Gasteiger, R. W. W. Hooft, K. Kristiansen, W. Kuipers, V. Laudet, L. Oliveira, M. Robinson, F. Rippmann, C. Sander, E. M. van der Wenden and A. P. IJzerman for stimulating discussions. We also acknowledge G. Valen, S. Sizemore and J. Weare for technical assistance. The NucleaRDB effort is supported by Organon and Lion BioSciences Ag. F.H. and F.E.C. acknowledge the NIH for support.

REFERENCES

1. Horn,F., Weare,J., Beukers,M.W., Horsch,S., Bairoch,A., Chen,W., Edvardsen,O., Campagne,F. and Vriend,G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.*, **26**, 275–279. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 346–349.
2. Horn,F., Mokrane,M., Weare,J. and Vriend,G. (2000) G protein-coupled receptors, or the power of data. In Suhai,S. (ed.), *Genomics and Proteomics: Functional and Computational Aspects*. Kluwer Academic/Plenum, New York, NY, pp. 191–214.
3. Baldwin,J.M., Schertler,G.F. and Unger,V.M. (1997) An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.*, **272**, 144–164.
4. Palczewski,K., Kumasaka,T., Hori,T., Behnke,C.A., Motoshima,H., Fox,B.A., Le Trong,I., Teller,D.C., Okada,T., Stenkamp,R.E., Yamamoto,M. and Miyano,M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, **289**, 739–745.
5. Gudermann,T., Nurnberg,B. and Schultz,G. (1995) Receptors and G proteins as primary components of transmembrane signal transduction. Part 1. G-protein-coupled receptors: structure and function. *J. Mol. Med.*, **73**, 51–63.
6. Gronemeyer,H. and Laudet,V. (1995) Transcription factors 3: nuclear receptors. *Protein Profile*, **2**, 1173–1308.
7. Mangelsdorf,D.J., Thummel,C., Beato,M., Herrlich,P., Schutz,G., Umesono,K., Blumberg,B., Kastner,P., Mark,M., Chambon,P. *et al.* (1995) The nuclear receptor superfamily: the second decade. *Cell*, **83**, 835–839.
8. Laudet,V. (1997) Evolution of the nuclear receptor superfamily: early diversification from an ancestral orphan receptor. *J. Mol. Endocrinol.*, **19**, 207–226.
9. Nuclear Receptors Comittee (1999) A unified nomenclature system for the nuclear receptor superfamily. *Cell*, **97**, 161–163.
10. Martinez,E., Moore,D.D., Keller,E., Pearce,D., Vanden Heuvel,J.P., Robinson,V., Gottlieb,B., MacDonald,P., Simons,S.,Jr, Sanchez,E. and Danielsen,M. (1998) The Nuclear Receptor Resource: a growing family. *Nucleic Acids Res.*, **26**, 239–241.
11. Oliveira,L., Paiva,A.C. and Vriend,G. (1993) A common motif in G protein-coupled seven transmembrane helix receptors. *J. Comput. Aided Mol. Des.*, **7**, 649–658.
12. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
13. Baker,W., van den Broek,A., Camon,E., Hingamp,P., Sterk,P., Stoesser,G. and Tuli,M.A. (2000) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **28**, 19–23. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 17–21.
14. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
15. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 214–218.
16. Hooft,R.W., Sander,C. and Vriend,G. (1996) Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins*, **26**, 363–376.
17. Laskowski,R.A., Hutchinson,E.G., Michie,A.D., Wallace,A.C., Jones,M.L. and Thornton,J.M. (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.*, **22**, 488–490.
18. Beukers,M.B., Kristiansen,K., IJzerman,A.P. and Edvardsen,O. (1999) TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data. *Trends Pharmacol. Sci.*, **20**, 475–477.
19. Gottlieb,B., Lehvaslaiho,H., Beitel,L.K., Lumbroso,R., Pinsky,L. and Trifiro,M. (1998) The Androgen Receptor Gene Mutations Database. *Nucleic Acids Res.*, **26**, 234–238.
20. Seeman,P. (1993) *Receptor Tables, vol.2: Drug dissociation constants for neuroreceptors and transporters*. SZ Research, Toronto, Canada.
21. Vriend,G. (1990*)* WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.*, **8**, 52–56, 29.
22. Oliveira,L., Paiva,A.C.M. and Vriend,G. (1995) Correlated mutations analysis of G protein α-chains to search for residues linked to binding. In Kaumaya,P.T.P. and Hodges,R.S. (eds), *Peptides: Chemistry, Structure and Biology*. Mayflower Scientific Ltd, Kingswinford, UK, pp. 408–409.
23. Kuipers,W., Oliveira,L., Paiva,A.C.M., Rippman,F., Sander,C., Vriend,G. and IJzerman,A.P. (1996) Sequence-function correlation in G protein-coupled receptors. In Findlay,J.B.C. (ed.), *Membrane protein models*. BIOS Scientific Publishers Ltd, Oxford, UK, pp. 27–45.
24. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
25. Walther,D. (1997) WebMol–a Java-based PDB viewer. *Trends Biochem. Sci.*, **22**, 274–275.
26. Campagne,F., Jestin,R., Reversat,J.L., Bernassau,J.M. and Maigret,B. (1999) Visualisation and integration of G protein-coupled receptor related information help the modelling: description and applications of the Viseur program. *J. Comput. Aided Mol. Des.*, **13**, 625–643.
27. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.