BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 23 2010, pages 3004–3005
doi:10.1093/bioinformatics/btq563

Databases and ontologies

Advance Access publication October 5, 2010

GPCRRD: G protein-coupled receptor spatial restraint database for 3D structure modeling and function annotation

Jian Zhang and Yang Zhang*

Center for Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109, USA

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: G protein-coupled receptors (GPCRs) comprise the largest family of integral membrane proteins. They are the most important class of drug targets. While there exist crystal structures for only a very few GPCR sequences, numerous experiments have been performed on GPCRs to identify the critical residues and motifs. GPCRRD database is designed to systematically collect all experimental restraints (including residue orientation, contact and distance maps) available from the literature and primary GPCR resources using an automated text mining algorithm combined with manual validation, with the purpose of assisting GPCR 3D structure modeling and function annotation. The current dataset contains thousands of spatial restraints from mutagenesis, disulfide mapping distances, electron cryo-microscopy and Fourier-transform infrared spectroscopy experiments.

Availability: http://zhanglab.ccmb.med.umich.edu/GPCRRD/

Contact: zhng@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 15, 2010; revised on September 14, 2010; accepted on September 29, 2010

1 INTRODUCTION

G-protein-coupled receptors (GPCRs) are the largest family of membrane proteins and mediate most cellular responses to hormones and neurotransmitters, as well as being responsible for vision, olfaction and taste (Rosenbaum *et al*., 2009). Many diseases involve the malfunction of these receptors, making them important drug targets. While knowledge of a protein's structure furnishes important information for understanding its function and drug design (Skolnick *et al*., 2000), the experimental determination of the 3D structure of GPCR membrane proteins has proved to be very difficult. Only four GPCR structures from human have been solved so far: β2-adrenergic (Rosenbaum *et al*., 2007), A2A adenosine (Jaakola *et al*., 2008), CXCR4 and Dopamine D3 (to be released). Fortunately, computer-based methods for predicting the 3D structure of a protein from its amino acid sequence have been increasingly successful as demonstrated by the recent CASP experiments (Moult *et al*., 2009). The structure models for all 907 registered GPCRs in the human genome were recently generated using a threading-assembly refinement method (TASSER; Zhang *et al*., 2006). About 820 GPCRs are anticipated to have correct topology and transmembrane helix arrangement. Nevertheless, great cautions are needed when utilizing the homology-based models for detail structural and functional annotations since helix kinks and extracellular loops are often different in different receptors. Modeling the subtle distinctions, which is essential for ligand docking and screening, remains a major challenge as highlighted by the recent blind GPCR Dock experiment (Michino *et al*., 2009).

While there exist crystal structures for only a very few GPCR sequences, numerous experiments have been performed on GPCRs to identify the critical residues and motifs. The information can be of important use to the structure and function modeling of the GPCR molecules. For example, the coherent activation and inactivation of residues in mutagenesis experiments usually indicate that the residues are spatially neighbors because they are binding to the common ligands (Becker *et al*., 2003; Du *et al*., 1997; Shacham *et al*., 2004; Shi and Javitch, 2002). The orientation of mutated functional residues is usually towards inside of the seven-helix bundle (Schushan *et al*., 2010). Thus, specific contacts or distance maps and residue orientations can be derived from the experimental data which can be used as restraints to guide the protein structure modeling simulations (Roy *et al*., 2010; Sali and Blundell, 1993; Zhang *et al*., 2003); this is especially helpful for the modeling of the structurally variant regions that cannot be directly transferred by homology inference (Paiva *et al*., 2006). GPCRRD is designed to systematically collect all available restraints derived from the experimental data scattered in the literature and GPCR-related databases using an automated text mining algorithm combined with manual validation, with the purpose of assisting GPCR 3D structure modeling and function annotation. The GPCRRD database is freely accessible. The database is updated automatically once every 2 weeks.

2 METHODS

GPCRRD consists of three steps of primary data collection, restraints derivation and data validation. A detailed description of the procedures is provided in the Supplementary Material. Here, we outline the major steps.

Experimental data are extracted from literature and other online databases using an automated procedure. Documents are first retrieved from the Medline database using the PubMed query system (Schuler, *et al*., 1996). Medline abstracts are used when full texts are not available. Pattern matching with regular expressions was used to identify point mutation data (see Equation S1). We then combined these data with those from the primary sources (GPCRDB; Horn *et al*., 2003, UniProt; Yip *et al*., 2008, EMBL; Kanz *et al*., 2005, TinyGRAP; Beukers *et al*., 1999) and the redundant entries were removed.

*To whom correspondence should be addressed.

Three different filters were applied to validate the experimental information. First, we applied a sequence filter to check whether the wild-type amino acids in the extracted point mutations are found at the indicated positions in the corresponding sequences. Secondly, we used a function filter to find the function related mutation. When the residue number of the mutation and the function-related words, such as ability, mediate, select, agonist, antagonist or binding, etc. occur in the same sentence and there is no 'not', we consider this mutation as functional mutation. Finally, we manually validated those data that do not belong to the two categories. Electron microscopy, neutron diffraction, Fourier-transform infrared spectroscopy (FTIR), disulfide bridge and X-ray data are collected and validated from the literature and from the original sources.

Three types of restraints are derived: residue orientation, side-chain contacts and distance map. To derive the residue orientation restraints, we first predict the transmembrane helices using TMHMM 2.0 (Krogh *et al*., 2001). Based on the assumption that the orientation of mutated functional residues is towards inside of the seven-helix bundle (Schushan *et al*., 2010; Shacham *et al*., 2004), the orientation restraint of mutation data can be obtained as showed in Supplementary Figures S2 and S3. If we define a plane $O_aO_bO_c$, which is perpendicular to the transmembrane helix TM_b and passing through the α-carbon atom of a query residue, we can have three points O_a, O_b and O_c, which are the intersections of plane $O_aO_bO_c$ and three axis of the neighboring transmembrane helices. These three points provide graphical representation of the lower and upper limits of the orientation restraint. The restraints of FTIR data can be generated in the same way.

For the disulfide bridge, we generated distance restraints according to geometry of disulfide bond. The 2D electron density maps of electron microscopy and neutron diffraction can be converted into 2D position restraints in the membrane surface plane. To generate residue contact restraints, we mainly consider the pair-wise mutagenesis and the agonist/antagonist binding data since the side-chains of residues binding to small common ligand are usually in a contact distance (Becker *et al*., 2003; Du *et al*., 1997; Shi and Javitch, 2002). Most of these data collections require manual checking and reading of the primary literatures.

3 RESULTS

The GPCRRD currently contains 10 electron microscopy, 2 neutron diffraction, 5588 functional site-directed mutagenesis, 16 FTIR, 38 disulfide bridge and 15 X-ray data. The Search page allows users to specify GPCRs of interest. A basic query system is available to search GPCRRD entries using SWISS-PROT identifiers or protein names. Another way to access the data is to browse lists of all the data in the database. An illustration plot, which is generated with the Graph Visualization Software-Graphviz, is used to represent and combine GPCR sequence and experimental data information. For each GPCR, a pointer is available to access remote information. This is done automatically by reading the SWISS-PROT entries and querying other remote resources. For each entry, cross-links to the original articles and other resources e.g. GPCRDB; Horn *et al*., 2003 and GPCR-OKB; Khelashvili *et al*., 2010 are provided.

4 CONCLUSION

The main purpose of GPCRRD is to automatically retrieve and update experimental information for GPCR 3D structure modeling and function annotation. Although several GPCR databases have been developed, useful information for GPCR 3D modeling cannot be directly obtained from these sources. For example, GPCRDB has general sequence and family data of GPCRs and TinyGRAP collects mutations but both do not provide the residue-level structural information. GPCRRD is to our knowledge the first structure-oriented database that systematically collects GPCR spatial restraints from primary experimental resources assisted by manual curation. The use of the GPCRRD to guide I-TASSER (Roy *et al*., 2010) for high-resolution GPCR structure and function modeling is under progress.

Funding: Alfred P. Sloan Foundation, National Science Foundation Career Award (DBI 1027394); National Institute of General Medical Sciences (GM083107, GM084222).

Conflict of Interest: none declared.

REFERENCES

Becker,O.M. *et al*. (2003) Modeling the 3D structure of GPCRs: advances and application to drug discovery. *Curr. Opin. Drug Disc.*, **6**, 353–361.

Beukers,M.W. *et al*. (1999) TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data. *Trends Pharmacol. Sci.*, **20**, 475–477.

Du,P. *et al*. (1997) Modeling the G-protein-coupled neuropeptide Y Y1 receptor agonist and antagonist binding sites. *Protein Eng.*, **10**, 109–117.

Horn,F. *et al*. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.,*, **31**, 294–297.

Jaakola,V.P. *et al*. (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science*, **322**, 1211–1217.

Kanz,C. *et al*. (2005) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **33**, D29–D33.

Khelashvili,G. (2010) GPCR-OKB: the G protein coupled receptor oligomer knowledge base. *Bioinformatics*, **26**, 1804–1805.

Krogh,A. *et al*. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Michino,M. *et al*. (2009) Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat. Rev. Drug Discov.*, **8**, 455–463.

Moult,J. *et al*. (2009) Critical assessment of methods of protein structure prediction-Round VIII. *Proteins*, **77**, 1–4.

Paiva,A.C. *et al*. (2006) Modeling GPCRs. *Ernst Schering Found. Symp. Proc.*, **2**, 23–47.

Rosenbaum,D.M. *et al*. (2007) GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science*, **318**, 1266–1273.

Rosenbaum,D.M. *et al*. (2009) The structure and function of G-protein-coupled receptors. *Nature*, **459**, 356–363.

Roy,A. *et al*. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.*, **5**, 725–738.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Schuler,G.D. *et al*. (1996) Entrez: Molecular biology database and retrieval system. *Comput. Methods Macromol. Seq. Anal.*, **266**, 141–162.

Schushan,M. *et al*. (2010) Cα-trace model of the transmembrane domain of human copper transporter 1, motion and functional implications. *Proc. Natl Acad. Sci.*, **107**, 10908–10913.

Shacham,S. *et al*. (2004) PREDICT modeling and in-silico screening for G-protein coupled receptors. *Proteins*, **57**, 51–86.

Shi,L. and Javitch,J.A. (2002) The binding site of aminergic G protein-coupled receptors: The transmembrane segments and second extracellular loop. *Annu. Rev. Pharmacol.*, **42**, 437–467.

Skolnick,J. *et al*. (2000) Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.*, **18**, 283–287.

Yip,Y.L. *et al*. (2008) Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum. Mutat.*, **29**, 361–366.

Zhang,Y. *et al*. (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.*, **85**, 1145–1164.

Zhang,Y. *et al*. (2006) Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Comput. Biol.*, **2**, 88–99.