Published in final edited form as:

Biopolymers. 2011 January ; 95(1): 31–38. doi:10.1002/bip.21528.

Comparison Analysis of Primary Ligand Binding Sites in Seven-Helix Membrane Proteins

Vagmita Pabuwat¹ and Zhijun Li^{1,2,*}

¹Department of Chemistry & Biochemistry, University of the Sciences in Philadelphia, Philadelphia, PA 19104, USA

²Institute for Translational Medicine and Therapeutics, University of the Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Seven-helix transmembrane proteins, including the G-protein coupled receptors, mediate a broad range of fundamental cellular activities through binding to a wide range of ligands. Understanding the structural basis for the ligand-binding selectivity of these proteins is of significance to their structure-based drug design. Comparison analysis of proteins' ligand binding sites provides a useful way to study their structure-activity relationships. Various computational methods have been developed for the binding site comparison of soluble proteins. In this work, we applied this approach to the analysis of the primary ligand-binding sites of 92 seven-helix transmembrane proteins. Results of the studies confirmed that the binding site of bacterial rhodopsins is indeed different from all G-protein coupled receptors. In the latter group, further comparison of the binding sites indicated a group of residues that could be responsible for ligand-binding selectivity and important for structure-based drug design. Further, unexpected binding site dissimilarities were observed among adrenergic and adenosine receptors, suggesting that the percentage of the overall sequence identity between a target protein and a template protein alone is not sufficient for selecting the best template for homology modeling of seven-helix membrane proteins. These results provided novel insight into the structural basis of ligand-binding selectivity of seven-helix membrane proteins and are of practical use to the computational modeling of these proteins.

Keywords

seven-helix membrane protein; GPCR; binding site; comparison analysis; cluster analysis

Introduction

Seven-helix membrane proteins, including bacterial rhodopsins and eukaryotic G-protein coupled receptors (GPCRs), are implicated in a wide range of cellular functions ¹. The GPCR superfamily comprises more than 800 proteins in human ² and plays a pivotal role in mediating signal transduction from outside to inside the cell. Further, GPCRs account for the targets of approximately 27% of drugs currently available in market ³ and remain an important one for future drug discovery ⁴. All seven-helix membrane proteins are believed to share a common topology of seven transmembrane α -helices with an extracellular amino terminus ⁵ ⁶. Moreover, for bacterial rhodopsins and many therapeutically important GPCRs, a roughly common binding site is determined to be at the extracellular portion of

*To whom correspondence should be addressed. Address: Department of Chemistry & Biochemistry, Box 100, University of the Sciences in Philadelphia, Philadelphia, PA 19104, USA Phone: 215-596-7539, Fax: 215-895-1112, z.li@usp.edu.

their transmembrane domain 6⁷. Intriguingly, through such a common binding platform, these proteins, in particular GPCRs, can selectively bind and be activated by a widely diverse set of ligands. Hence, understanding the structural basis of their ligand-binding selectivity represents an important aspect of structure-function relationships of these proteins and is also crucial to structure-based approaches for developing drugs that selectively target a specific GPCR subtype⁸. This, however, remains as a challenge despite intensive research efforts.

A major difficulty is that most GPCRs are extremely difficult to crystallize and for many years, the experimental structure of a bacterial rhodopsin, which shares retinal, the common natural inverse agonist of rhodopsin GPCRs, had been adopted as the prototype for modeling studies of GPCR molecules⁹. Experimental breakthroughs in recent years have resulted in the availability of more than twenty X-ray structures for members of GPCRs, including the opsin receptors^{7, 10}, the β_1 ¹¹ and β_2 adrenergic receptor^{12, 13}, and the adenosine A2A receptor¹⁴. Detailed structure-based analyses, combined with sequence-based analyses, have provided novel insight into the mechanism of ligand-binding selectivity. For instances, the conservation pattern of amino acid residues at the ligand-binding site has long been studied in order to identify key residues for selectivity¹⁵. The second extracellular loop in the β_1 receptor was considered to play a key role in ligand-binding specificity¹¹.

Despite these progresses, predicting the ligand-binding selectivity of a GPCR of unknown structure remains elusive. Due to the paucity of available high-resolution structures, for many GPCRs, their homology models have to be constructed using structural templates of low sequence identity. Consequently, the quality of such models is uncertain and their application is often limited. A popular approach to address this involves first constructing the homology models of the target protein using all available template structures respectively, and subsequently evaluating these models by their binding affinity to known ligands through ligand docking techniques^{16–19}.

Complementary to sequence-based and structure-based approaches, comparing the ligand-binding sites of protein structures may help with the understanding of their ligand selectivity (Figure 1). Proteins with similar functional sites in terms of both shape and physico-chemical properties bind to similar ligands, even though their structures adopt different folds²⁰. The extent of the binding site similarity is also essential for selecting appropriate structural templates for homology modeling of GPCRs. Given the importance of such an analysis, a variety of computational tools have been developed to perform comparison analysis of ligand-binding sites^{21–22}. These tools have been applied to the functional classification of enzymes including kinases²³, to the detection of similarity across different protein families²⁴, and to the understanding of cross reactivity²⁵.

In this work, the comparison of the primary ligand-binding sites of seven-helix membrane proteins was investigated using Cavbase²⁶, a component of the Relibase+ package²⁷. Cavbase describes and compares protein binding sites based on both the geometrical and physico-chemical property²⁸. A structure dataset of 92 seven-helix membrane proteins was prepared and their primary ligand-binding sites were compared. The studies showed that the binding sites of bacterial rhodopsins are quite different from GPCRs including rhodopsin GPCRs. They also suggested a group of residues that could be responsible for ligand-binding selectivity and could be important for structure-based drug design for GPCRs. More important, unexpected binding site dissimilarities were observed, which suggested that the percentage of sequence identity between a target protein and a template protein alone is not sufficient for selecting the best template for homology modeling of seven-helix membrane proteins, particularly for the purpose of structure-based drug design. These results provided

novel insight into the ligand-binding selectivity of seven-helix membrane proteins and may help with the application of homology modeling techniques to this class of membrane proteins.

Materials and Methods

Seven-helix membrane protein dataset

A structure dataset of seven-helix membrane proteins was compiled by searching the Protein Data Bank (PDB) 29. First, ten sub-family names of seven-helix membrane proteins were identified from the online database of membrane protein (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html#Latest, version Nov. 24, 2009); Next, these names were adopted for keyword search of PDB and the hits were compared with the entries in the online database above to ensure that no entries were missing. Finally, all the hits were winnowed with the following criteria: (i) The structure was determined by X-ray crystallographic methods at a resolution of 4.0Å or better; (ii) The structure contains a ligand with C atoms, which bound to the extracellular portion of the transmembrane domain of the protein; (iii) The PDB file of the structure can be opened using Hermes, a visualization tool in the Relibase+ software package (Cambridge Crystallographic Data Centre, version 3.0.0) 26; and (iv) If the PDB file of a structure contains more than one model, the first one was selected unless otherwise specified.

The final dataset contained 92 structures of seven-helix membrane proteins (Table I). Among them, 77 were bacterial rhodopsins and 15 were GPCRs. Among the bacterial rhodopsins, 22 structures (Table II) were identified in the literature as the bacteriorhodopsin at different conformational states 30.

Comparison of ligand-binding sites

The binding site residues of the protein structures in the compiled dataset were defined as those within the 4.5Å of the ligand atoms. The 4.5Å distance cutoff has long been adopted as a simple and computationally fast way to define an interaction between the binding site residues and the ligand 35. For proteins and their bio-organic ligands, the van der Waals radii of their common atom types are typically less than or equal to 2.0Å. The 4.5 Å cutoff is thus more than the sum of the van der Waals radii of two such atoms. For bacteriorhodopsin structures, an additional cutoff of 6.0Å was also employed. The purpose of this increase was to ensure that considering additional residues as part of the binding site won't change the conclusion of the analyses.

Comparison between two binding sites was carried out using Cavbase 26. Cavbase represents a binding site using a set of pseudocenters defined as a point in 3D space. In addition, each pseudocenter is associated with one of the physicochemical properties such as hydrogen bond (HB) donor, HB acceptor, hydrophobic contact, etc. The pseudocenters were further mapped to the surface of the binding site. Thus, this condensed representation allows for efficient comparison of two binding sites in both shape and physicochemical properties.

To apply Cavbase for this study, all 92 structures from the dataset were loaded to Hermes and for each structure, pseudocenters were defined and mapped to the site surface as described above (Figure 1). An all-to-all comparison was then performed for all binding sites using the clique detection algorithm 26 and the similarity was measured by three built-in scoring functions, SF1, SF2 and SF3, respectively. These scoring functions measure the binding site similarity based on the percentage of the number of matched pseudocenters for the two binding sites, as detected by the clique algorithm. For each comparison between a query binding site and the binding sites of the other 91 proteins, the default parameters were

adopted with the exception for the *Number of Solutions to Save*, which was changed from 100 to 500 to ensure all comparisons were saved. This change was due to the fact that the software automatically detects all possible binding sites on the 91 proteins and compares them with the query binding site.

Binding site clustering and further analysis

The scores obtained through pairwise comparison above were converted into a matrix and were subsequently adopted as the input for the clustering algorithm CLUTO 31. CLUTO is a free clustering software from Karypis' lab (<http://glaros.dtc.umn.edu/gkhome/cluto/cluto>). Four different algorithms, rb, rbr, agglo and graph, were implemented in the software. The first three algorithms, together with the first scoring function SF1, were shown to generate optimal results for classification 28. For this work, the rb algorithm combined with SF1 was chosen for the final clustering analysis. To cluster the 92 structures in the dataset, we used the eight-way clustering since the number of protein families in our dataset was already known (Table I). Similarly, we used the six-way clustering for the 22 structures of bacteriorhodopsins at various conformational states. For the 15 structures of GPCRs, the three-way clustering was also adopted since the structures represent three GPCR families.

To further analyze the results of the binding site comparison, three structure pairs of GPCRs were selected. For each structure, its transmembrane helical boundaries were identified based on the definition in the PDBTM database³⁶. For each transmembrane helix, multiple sequence alignments of the six helical sequences of the three selected structure pairs were generated using the AMPS package³⁷. Binding site residues for each structure pair that either matched or non-matched with each other were highlighted in the multiple sequence alignment with ALSCRIPT³⁸. Whenever necessary, additional residues were included for the helical sequences so that for each transmembrane helix, the six helical sequences have the same length in the alignment.

Unexpected binding site similarity and dissimilarity in GPCRs

Pairwise sequence alignment for all 15 GPCR proteins was performed using the EMBOSS pairwise alignment algorithm at the website of the European Bioinformatics Institute (<http://www.ebi.ac.uk/Tools/emboss/align/>) with the default settings. For all pairs, the relationship between the percentage of their sequence identity and their binding site similarity score (SF1) was fitted to a linear function. Unexpected binding site similarities and dissimilarities between pairs were identified.

Results

To compare the primary ligand-binding sites in seven-helix membrane proteins, the computational approach included several steps: (i) Compile an X-ray structure dataset of those membrane proteins, including 22 structures of bacteriorhodopsin at different conformational states and 15 structures of GPCRs; (ii) Identify and compare the primary ligand binding sites of all structures in the dataset; and (iii) Cluster the binding sites and analyze the results.

Overall binding site similarity of seven-helix membrane proteins

Based on the binding site comparison, the 92 seven-helix membrane protein structures in the dataset fell into approximately four groups (Figure 2). The first group included the five families belonging to the bacterial rhodopsin class (Table I). Proteins within this group all displayed much higher similarity with each other than with members of other groups. The other three groups included the adrenergic receptors (β_1 and β_2), the adenosine receptor, and the opsin receptors (bovine and squid rhodopsin). Similarly, members of each GPCR group

all displayed much higher similarity with each other than with members of other groups. This was shown to be true with either 4.5Å or 6.0Å cutoff for defining binding site residues. The results were consistent with the current classification of these proteins. They also suggested that consistent with their low sequence identity (<20%), there is significant difference between the ligand-binding sites of bacterial rhodopsins and rhodopsin GPCRs, despite the fact that they both bind to the same ligand of retinal.

Binding site comparison of bacteriorhodopsins at different conformational states

Bacteriorhodopsins undergo a cycle of five intermediate states upon illumination (Table II), starting from the ground or unilluminated state, then going through a series of intermediate states (K-L-M-N-O), finally returning to the ground state 30–32. The mechanism of the photon transport upon illumination and the conformational changes in various states remain unclear. Comparison of the binding sites of bacteriorhodopsins in these different conformational states could be of help to the understanding of the structural basis for their activation.

We compared the binding sites of 22 bacteriorhodopsin structures (Table II), representing the ground and all five intermediate states. Based on the six-way clustering, the ground state, the K state and almost all the structures of the M state were clearly clustered together (Figure 3). The N state was clustered separately. However, overall the binding sites of structures of various states retained similarity. This is consistent with the current view that the conformational changes upon illumination are global and not localized near the ligand 30.

Binding site comparison of GPCRs

GPCR structures included in the dataset belong to three families (Table I). Accordingly, in the binding site comparison analysis, members from the same family showed higher similarity scores with each other and were clustered together with no exception (Figure 4). This was true even for the β_1 and β_2 adrenergic receptors, suggesting they contain a certain level of similarity at their binding sites.

To understand the structural basis of the ligand-binding selectivity of different sub-families of GPCRs, the matching and non-matching pseudocenters and their corresponding binding site residues were compared. In most cases, binding site residues that matched well with each other are conserved within an individual family or even the entire GPCR super-family (Figure 5). On the other hand, conserved binding site residues may not match with each other.

Interestingly, non-matching binding site residues from different families, which are assumed to be responsible for ligand-binding selectivity, belong to different transmembrane helices. For bovine and squid rhodopsins, those residues were primarily on the third helix; While for β_2 adrenergic and the adenosine receptors, they were scattered at the helices three, six and seven. In addition, for the β_1 and β_2 adrenergic receptors, essentially all binding site residues matched with each other, consistent with the finding that the second extracellular loop in the β_1 receptor was considered to play a key role in ligand-binding specificity 11.

Unexpected similarity and dissimilarity in the binding site of GPCRs

Proteins with the high sequence similarity are generally assumed to have high structural similarity 33. Consistently, a clear linear relationship between the percentage of the sequence identity of GPCR pairs and their binding site similarity score was observed (Figure 6). However, detailed analysis of results of binding site comparison among GPCRs showed that unexpected similarity and dissimilarity existed for the binding sites of GPCRs. For example, the β_2 adrenergic receptor shared the sequence identity of >50% with the A2A

adenosine receptor, but their binding site similarity score was as low as the scores between the bovine receptors and the A_{2A} adenosine receptor (Figure 6). On the other hand, the human β_1 and β_2 adrenergic receptor had the sequence identity of ~35% with each other, their binding sites were nearly identical (Figure 6). Further examination of the sequence alignment indicated that this unexpected disparity was primarily due to the non-transmembrane sub-sequences (Table III) as those non-transmembrane regions probably evolve at different rates from the transmembrane regions.

Discussion

Understanding the structural basis for the ligand-binding selectivity of seven-helix membrane proteins, particularly GPCRs, is essential for elucidating the structure-function relationships of these biologically and therapeutically important proteins as well as their structure-based drug-discovery efforts 8-11. This challenge has been intensively studied through computational approaches including sequence analysis and direct structure analysis 11-15. Comparison analysis of binding sites of proteins complements the sequence-based and fold-based analysis in protein function studies 21-22. In this work, this approach was applied to the studies of X-ray structures of seven-helix membrane proteins.

A number of tools have been developed to carry out comparison analysis of protein functional sites 21. For this study, we chose the Cavbase package developed at Dr. Kebe's lab after working with several similar tools. Cavbase represents the geometrical and physico-chemical properties of a binding site using a set of pseudocenters 26. For a binding site comparison, the pseudocenters from both sites are superimposed and the percentage of the shared pseudocenters is calculated. This algorithm runs very fast and has proven to be quite accurate in several studies, including enzyme function-based classification 28.

Applying the binding site comparison to a dataset of seven-helix membrane proteins has resulted in some useful conclusions. First, our results showed that the clustering of seven-helix membrane proteins based on their binding site similarity works well and is very consistent with knowledge-based classification. Both bacterial rhodopsins and eukaryotic GPCRs are classified separately and within each class, proteins considered in the same family were indeed clustered together (Figure 2). Further, it provides a rational explanation for the cross-reactivity of the β_1 and β_2 adrenergic receptors as well as the bovine and the squid rhodopsins, since proteins from the two sub-families of the same family indeed showed relatively high binding site similarity (Figure 4). In addition, the clustering of various conformational states of bacteriorhodopsins supported the literatures and experimental results 30.

Surprisingly, the binding site of bacterial rhodopsins showed little similarity to rhodopsin GPCRs despite the fact that both family members bind to the same inverse agonist of retinal (Figure 2). Further examination of these structures suggested one possible reason that the retinal ligand adopted different conformations while bound to different binding pockets. On the other hand, the sequence identity between bacterial rhodopsins and rhodopsin GPCRs is quite low (<20%). Therefore, both sequence-based and binding site-based similarity analyses showed that structures of bacterial rhodopsins were not good templates for modeling of GPCRs.

Further examination of the matching and non-matching binding site pseudocenters and residues has resulted in some interesting observations (Figure 5). In most cases, binding site residues that are conserved through the individual family or even the entire GPCR super-family tend to match well between members of two sub-families. On the other hand, conserved binding site residues may not match with each other. This suggests that limitation

of sequence-based conservation analysis for identifying residues that is responsible for ligand-binding selectivity. For different families, the locations of non-matched binding site residues also varied. For bovine and squid rhodopsins, those residues were primarily on the third helix; While for β_2 adrenergic and the adenosine receptors, they were scattered at the helices three, six and seven. In addition, for the β_1 and β_2 adrenergic receptors, essentially all binding site residues matched with each other. This is consistent with the finding that the second extracellular loop in the β_1 receptor was considered to play a key role in ligand-binding specificity 11. These observations have provided useful ideas for structure-based drug design. For instance, to increase or decrease the binding selectivity of a lead compound targeting a specific sub-type of GPCRs, the focus should be on the chemical structure of the lead compound that is close to the non-matched pseudocenters. By adding or deleting a functional group, the interaction between the compound and the corresponding residues could be changed.

Perhaps most interestingly, our studies have uncovered unexpected bindings site similarity and dissimilarity. For instance, although the β_2 adrenergic receptor (PDB ID: 2RH1) and the adenosine receptor (PDB ID: 3EML) share 53% of sequence identity, their binding sites are significantly different (Figure 5). Clearly, neither the bovine rhodopsin nor the adrenergic receptor is an ideal template for modeling of the adenosine receptor 19, though templates with more than 50% of sequence similarity to a target protein are generally considered to be able to generate a good-quality homology model 33. This result demonstrated that relatively high overall sequence similarity does not always signify high binding site similarity, suggesting the importance of direct comparison of the binding sites of seven-helix membrane proteins. As only a small number of X-ray structures is currently reported, for many GPCRs, their homology models can only be constructed based on templates with low sequence identity 34. The work presented here again emphasizes the importance of model refinement and verification in the current stage of constructing homology models for studies of seven-helix membrane proteins.

Acknowledgments

The authors thank Dr. Karypis's lab for the CLUTO software freely available to us. We thank the CCDC staffs for their technical support and Dr. Michael F. Bruist at University of the Sciences in Philadelphia for comments on the manuscript. This work was supported by the NIH grant R15-GM084404.

References

1. Fredriksson R, Schioth HB. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol. Pharmacol.* 2005; 67:1414–1425. [PubMed: 15687224]
2. Foord SM, Bonner TI, Neubig RR, Rosser EM, Pin JP, Davenport AP, Spedding M, Harmar AJ. International Union of Pharmacology. XLVI. G protein-coupled receptor list. *Pharmacol. Rev.* 2005; 57:279–288. [PubMed: 15914470]
3. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? *Nat. Rev. Drug Discov.* 2006; 5:993–996. [PubMed: 17139284]
4. Lagerstrom MC, Schioth HB. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat. Rev. Drug Discov.* 2008; 7:339–357. [PubMed: 18382464]
5. Baldwin JM, Schertler GF, Unger VM. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.* 1997; 272:144–164. [PubMed: 9299344]
6. Ballesteros JA, Shi L, Javitch JA. Structural mimicry in G protein-coupled receptors: implications of the high-resolution structure of rhodopsin for structure-function analysis of rhodopsin-like receptors. *Mol. Pharmacol.* 2001; 60:1–19. [PubMed: 11408595]

7. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*. 2000; 289:739–745. [PubMed: 10926528]
8. Costanzi S. On the applicability of GPCR homology models to computer-aided drug discovery: a comparison between in silico and crystal structures of the beta2-adrenergic receptor. *J. Med. Chem.* 2008; 51:2907–2914. [PubMed: 18442228]
9. Bikker JA, Trumpp-Kallmeyer S, Humblet C. G-Protein coupled receptors: models, mutagenesis, and drug design. *J. Med. Chem.* 1998; 41:2911–2927. [PubMed: 9685229]
10. Murakami M, Kouyama T. Crystal structure of squid rhodopsin. *Nature*. 2008; 453:363–367. [PubMed: 18480818]
11. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AG, Tate CG, Schertler GF. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature*. 2008; 454:486–491. [PubMed: 18594507]
12. Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Yao XJ, Weis WI, Stevens RC, Kobilka BK. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science*. 2007; 318:1266–1273. [PubMed: 17962519]
13. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK, Stevens RC. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science*. 2007; 318:1258–1265. [PubMed: 17962520]
14. Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien EY, Lane JR, Ijzerman AP, Stevens RC. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science*. 2008; 322:1211–1217. [PubMed: 18832607]
15. Kontoyianni M, DeWeese C, Penzotti JE, Lybrand TP. Three-dimensional models for agonist and antagonist complexes with beta 2 adrenergic receptor. *J. Med. Chem.* 1996; 39:4406–4420. [PubMed: 8893835]
16. Yuzlenko O, Kiec-Kononowicz K. Molecular modeling of A1 and A2A adenosine receptors: comparison of rhodopsin- and beta2-adrenergic-based homology models through the docking studies. *J. Comput. Chem.* 2009; 30:14–32. [PubMed: 18496794]
17. Sherbiny FF, Schiedel AC, Maass A, Muller CE. Homology modelling of the human adenosine A(2B) receptor based on X-ray structures of bovine rhodopsin, the beta(2)-adrenergic receptor and the human adenosine A (2A) receptor. *J. Comput. Aided Mol. Des.* 2009; 23(11):807–828. [PubMed: 19757091]
18. Mobarec JC, Sanchez R, Filizola M. Modern Homology Modeling of G-Protein Coupled Receptors: Which Structural Template to Use? *J. Med. Chem.* 2009; 52(16):5207–5216. [PubMed: 19627087]
19. Katritch V, Rueda M, Lam PC, Yeager M, Abagyan R. GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex. *Proteins*. 2010; 78:197–211. [PubMed: 20063437]
20. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: approaches and limitations. *Nat. Struct. Biol.* 2000; 7 Suppl:991–994. [PubMed: 11104008]
21. Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. U. S. A.* 2008; 105:5441–5446. [PubMed: 18385384]
22. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.* 2007; 368:283–301. [PubMed: 17337005]
23. Kuhn D, Weskamp N, Hullermeier E, Klebe G. Functional classification of protein kinase binding sites using Cavbase. *ChemMedChem*. 2007; 2:1432–1447. [PubMed: 17694525]
24. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A, Klebe G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* 2004; 47:550–557. [PubMed: 14736236]

25. Kinnings SL, Jackson RM. Binding site similarity analysis for the functional classification of the protein kinase family. *J. Chem. Inf. Model.* 2009; 49:318–329. [PubMed: 19434833]
26. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* 2002; 323:387–406. [PubMed: 12381328]
27. Hendlich M, Bergner A, Gunther J, Klebe G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.* 2003; 326:607–620. [PubMed: 12559926]
28. Kuhn D, Weskamp N, Schmitt S, Hullermeier E, Klebe G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* 2006; 359:1023–1044. [PubMed: 16697007]
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res.* 2000; 28:235–242. [PubMed: 10592235]
30. Hirai T, Subramaniam S. Protein conformational changes in the bacteriorhodopsin photocycle: comparison of findings from electron and X-ray crystallographic analyses. *PLoS One.* 2009; 4:e5769. [PubMed: 19488399]
31. Zhao Y, Karypis G. Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery.* 2005; 10:141–168.
32. Hirai T, Subramaniam S, Lanyi JK. Structural snapshots of conformational changes in a seven-helix membrane protein: lessons from bacteriorhodopsin. *Curr. Opin. Struct. Biol.* 2009; 19:433–439. [PubMed: 19643594]
33. Baker D, Sali A. Protein structure prediction and structural genomics. *Science.* 2001; 294:93–96. [PubMed: 11588250]
34. Michino M, Abola E. GPCR Dock 2008 participants, Brooks C. L., 3rd, Dixon J. S., Moulton J. & Stevens R. C. (2009) Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat. Rev. Drug Discov.* 2009; 8:455–463. [PubMed: 19461661]
35. Shoichet BK, Bodian DL, Kuntz ID. Molecular docking using shape descriptors. *J. Comp. Chem.* 1992; 13:380–397.
36. Tusnady GE, Dosztanyi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.* 2005; 33(Database issue):D275–D278. [PubMed: 15608195]
37. Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* 1993; 9:745–756. [PubMed: 8143162]
38. Barton GJ. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.* 1993; 6:37–40. [PubMed: 8433969]

Fig. 1.
Illustration of three computational approaches for studying protein sequence-structure-function relationships.

Fig. 2. Clustering analysis of the binding sites of the seven-helix membrane protein dataset. The rb clustering algorithm and scoring function SF1 were used. The mutual similarity of the binding sites, computed by the scoring function SF1, was indicated by the intensity of the red color (dark red, represented similarity; white, no similarity). The GPCR class (top four clusters) was clearly separated from the bacterial rhodopsin class (bottom four clusters).

Fig. 3. Clustering analysis of the 22 bacteriorhodopsin structures of various conformational states. The rb clustering algorithm and scoring function SF1 were used. The mutual similarity of the binding sites, computed by the scoring function SF1, was indicated by the intensity of the red color (dark red, represented similarity; white, no similarity). G represents the unilluminated or ground state. The conformational state of each structure was listed in Table II. Conformational states for each cluster were labeled at the bottom.

Fig. 4.

Clustering analysis of the GPCR structures included in the dataset. Only five structures of different sub-families were presented to allow effective comparison. The rb clustering algorithm and scoring function SF1 were used. The mutual similarity of the binding sites, computed by the scoring function SF1, was indicated by the intensity of the red color (dark red, represented similarity; white, no similarity).

2Z73 VYYS L G I F I G I C G I I G C G G N G I V I Y L F T
 1U19 Q F S M L A A Y M F L L I M L G F P I N F L T L Y V T V

2RH1 W V V G M G I V M S L I V L A I V F G N V L V I T A I A
 2VT4 W E A G M S L L M A L V V L L I V A G N V L V I A A I G

2RH1 W V V G M G I V M S L I V L A I V F G N V L V I T A I A
 3EML S S V Y I T V E L A I A V L A I L G N V L V C W A V W

TM1

M F I I N L A F S D F T F S L V N G F P L M T I S C
 Y I L L N L A V A D L F M V F G G F T T T L Y T S

Y F I T S L A C A D L V M G L A V V P F G A A H I L
 L F I T S L A C A D L V V G L L V V P F G A T L V V

Y F I T S L A C A D L V M G L A V V P F G A A H I L
 Y F V V S L A A A D I A V G V L A I P F A I T I S T

TM2

2Z73 F A A C K V Y G F I G G I F G F M S I M T M A M I S I D R Y
 1U19 L T G C N L E G F F A T L G G E I A L W S L V V L A I E R Y

2RH1 F G N F W C E F W T S I D V L C V T A S I E T L C V I A V D R Y
 2VT4 W G S F L C E L W T S L D V L C V T A S I E T L C V I A I D R Y

2RH1 F G N F W C E F W T S I D V L C V T A S I E T L C V I A V D R Y
 3EML G C L F I A C F V L V L T Q S S I F S L L A I A I D R Y

TM3

A F I M I I F V W L W S V L W A I G P I F G W G A Y
 A I M G V A F T W M A L A C A A P P L V G W S E

A R V I I L M V W I V S G L T S F L P I Q M H W Y R
 A K V I I C T V W A I S A L V S F L P I M M H W W R

A R V I I L M V W I V S G L T S F L P I Q M H W Y R
 A K G I I A I C W V L S F A I G L T P M L G W N N

TM4

2Z73 D S T T R S N I L C M F I L G F F G P I L I I F F C Y F
 1U19 E T N N E S F V I Y M F V V H F I I P L I V I F F C Y G

2RH1 F F T N Q A Y A I A S S I V S F Y V P L V I M V F V Y S
 2VT4 F V T N R A Y A I A S S I I S F Y I P L L I M I F V A L

2RH1 F F T N Q A Y A I A S S I V S F Y V P L V I M V F V Y S
 3EML V V P M N Y M V Y F N F A C V L V P L L M L G V Y L

TM5

S I V I V S Q F L L S W S P Y A V V A L L A Q F G
 V I I M V I A F L I C W L P Y A G V A F Y I F T H

L G I I M G T F T L C W L P F F I V N I V H V I Q
 L G I I M G V F T L C W L P F F L V N I V N V F N

L G I I M G T F T L C W L P F F I V N I V H V I Q
 L A I I V G L F A L C W L P L H I I N C F T F F C

TM6

2Z73 V T P Y A A Q L P V M F A K A S A I H N P M I Y S V S
 1U19 F G P I F M T I P A F F A K T S A V Y N P V I Y I M M

2RH1 I R K E V Y I L L N W I G Y V N S G F N P L I Y C R S
 2VT4 V P D W L F V A F N W L G Y A N S A M N P I I Y C R S

2RH1 I R K E V Y I L L N W I G Y V N S G F N P L I Y C R S
 3EML A P L W L M Y L A I V L S H T N S V V N P F I Y A Y R

TM7

Fig. 5.
 Multiple sequence alignments for the transmembrane regions (TM-1-TM-7) of representative GPCR pairs. Blackened blocks indicated matched positions in the binding site comparison using Cavbase and light gray blocks highlighted non-matched positions.

Fig. 6.

Correlation between the binding site similarity *and* the percentage of the sequence identity for each GPCR pair. For the 15 GPCR structures, totally 105 pairs were compared. a, 2VT4-2RH1; b, 2VT4-3D4S; c, 3EML-2RH1; and d, 3EML-3D4S.

Table I

Seven-helix transmembrane protein structures used in the analysis.

<i>Class</i>	<i>Protein Family</i>	<i>Species Name</i>	<i>PDB ID</i>
Bacterial Rhodopsin	Bacteriorhodopsin	<i>Halobacterium salinarum</i>	1AP9, 1AT9, 1BM1, 1BRR, 1BRX, 1C3W, 1C8R, 1C8S, 1CWQ, 1DZE, 1E12, 1E0P, 1F4Z, 1F50, 1FBB, 1FBK, 1IW6, 1IW9, 1IXF, 1JV6, 1JV7, 1KG8, 1KG9, 1KGB, 1KME, 1M0K, 1M0L, 1M0M, 1MGY, 1O0A, 1P8H, 1P8I, 1P8U, 1PXR, 1PXS, 1PY6, 1Q5I, 1Q5J, 1QHJ, 1QKO, 1QKP, 1QM8, 1S51, 1S52, 1S53, 1S54, 1S8J, 1S8L, 1TN0, 1TN5, 1UCQ, 1VJM, 1X0I, 1X0K, 1X0S, 1XJI, 2AT9, 2BRD, 2I20, 2I21, 2I1X, 2NTU, 2NTW, 2ZFE, 3COC, 3COD
	Sensory rhodopsin I	<i>Natronomonas pharaonis</i>	1XIO
	Sensory rhodopsin II	<i>Nostoc sp</i>	1GU8, 1GUE, 1H2S, 1H68, 1JGJ
	Xanthorhodopsin	<i>Salinibacter ruber</i>	3DDL
	Archaeorhodopsin	<i>Halobacter sp</i>	1UAZ, 1VGO, 2EI4, 2Z55
GPCR	Opsin receptor (bovin and squid)	<i>Bos taurus</i>	1F88, 1GZM, 1HZX, 1L9H, 1U19, 2G87, 2HPY, 2I35, 3C9M
		<i>Todarodes pacificus</i>	2Z73, 2ZIY
	Adrenergic receptor (β_1 and β_2)	<i>Meleagris galopavo</i>	2VT4
		<i>Homo sapiens</i>	2RH1, 3D4S
	Adenosine receptor (A2A)	<i>Homo sapiens</i>	3EML

Table II

Bacteriorhodopsin structures of various conformational states used for analysis.

<i>State</i>	<i>PDB ID</i>	<i>Sequence</i>
<i>Unilluminated</i>	1M0L	WT
	1C3W	WT
	1IW6	WT
	1KGB	WT
	1QHJ	WT
<i>K</i>	1QKP	WT
	1M0K:2	WT
	1IXF	WT
<i>L</i>	1E0P:B	ET
	1O0A	WT
	1UCQ	WT
	1VJM:2	WT
<i>M</i>	1KG8	WT
	1M0M:2	WT
	1P8H	WT
	1CWQ:B	WT
	1IW9	WT
	1F4Z	E204Q
	1C8S	D96N
<i>N</i>	1P8U	V49A
<i>O</i>	1JV7	D85S
	1X0I	WT

Table III

Percentage of sequence identity for selected GPCR structures.

<i>Pairs</i>	<i>Percentage of Sequence Identity (%)</i>	
	<i>Entire Sequence</i>	<i>Transmembrane Region</i>
β_2 adrenergic receptor (PDB ID: 2RH1) vs. β_1 adrenergic receptor (PDB ID: 2VT4) β_2 adrenergic receptor (PDB ID: 2RH1) vs. A2A adenosine receptor (PDB ID: 3EML)	36.3 53.1	57.1 32.5