Structure Modeling of All Identified G Protein–Coupled Receptors in the Human Genome

Yang Zhang, Mark E. DeVries, Jeffrey Skolnick*

Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York, United States of America

G protein–coupled receptors (GPCRs), encoded by about 5% of human genes, comprise the largest family of integral membrane proteins and act as cell surface receptors responsible for the transduction of endogenous signal into a cellular response. Although tertiary structural information is crucial for function annotation and drug design, there are few experimentally determined GPCR structures. To address this issue, we employ the recently developed threading assembly refinement (TASSER) method to generate structure predictions for all 907 putative GPCRs in the human genome. Unlike traditional homology modeling approaches, TASSER modeling does not require solved homologous template structures; moreover, it often refines the structures closer to native. These features are essential for the comprehensive modeling of all human GPCRs when close homologous templates are absent. Based on a benchmarked confidence score, approximately 820 predicted models should have the correct folds. The majority of GPCR models share the characteristic seven-transmembrane helix topology, but 45 ORFs are predicted to have different structures. This is due to GPCR fragments that are predominantly from extracellular or intracellular domains as well as database annotation errors. Our preliminary validation includes the automated modeling of bovine rhodopsin, the only solved GPCR in the Protein Data Bank. With homologous templates excluded, the final model built by TASSER has a global C_α root-mean-squared deviation from native of 4.6 Å, with a root-mean-squared deviation in the transmembrane helix region of 2.1 Å. Models of several representative GPCRs are compared with mutagenesis and affinity labeling data, and consistent agreement is demonstrated. Structure clustering of the predicted models shows that GPCRs with similar structures tend to belong to a similar functional class even when their sequences are diverse. These results demonstrate the usefulness and robustness of the in silico models for GPCR functional analysis. All predicted GPCR models are freely available for noncommercial users on our Web site (http://www.bioinformatics.buffalo.edu/GPCR).

Citation: Zhang Y, DeVries ME, Skolnick J (2006) Structure modeling of all identified G protein–coupled receptors in the human genome. PLoS Comput Biol 2(2): e13.

Introduction

G protein–coupled receptors (GPCRs) are integral membrane proteins embedded in the cell surface that transmit signals to cells in response to stimuli such as light, Ca^{2+}, odorants, amino acids, nucleotides, peptides, or proteins and mediate many physiological functions through their interaction with heterotrimeric G proteins [1,2]. Many diseases involve the malfunction of these receptors, making them important drug targets. In human, the estimated number of GPCRs is approximately 948 [3], corresponding to about 5% of the total number of human genes [4]. However, more than 45% of all modern drugs target GPCRs; these represent around 25% of the 100 top-selling drugs worldwide [2,5].

While knowledge of a protein's structure furnishes important information for understanding its function and for drug design [6], progress in solving GPCR structures has been slow [7]. Nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography are the two major techniques used to determine protein structures. NMR spectroscopy has the advantages that the protein does not need to be crystallized and dynamical information can be extracted. However, high concentrations of dissolved proteins are needed; and as yet no complete GPCR structure has been solved by the method. X-ray crystallography can provide very precise atomic information for globular proteins, but GPCRs are extremely difficult to crystallize. In fact, only a single GPCR, bovine rhodopsin (RH) from the rod outer segment membrane, has been solved [8]. It is unlikely that a significant number of high-resolution GPCR structures will be experimentally solved in the very near future. This situation limits the use of structure-based approaches for drug design and restricts research into the mechanisms that control ligand binding to GPCRs, activation and regulation of GPCRs, and signal transduction mediated by GPCRs [9].

Fortunately, as demonstrated by the recent CASP experiments [10], computer-based methods for deducing the three-dimensional structure of a protein from its amino acid sequence have been increasingly successful. Among the three types of structure prediction algorithms—homology model-

Editor: Diana Murray, Cornell University, United States of America

Received October 19, 2005; **Accepted** January 11, 2005; **Published** February 17, 2006

DOI: 10.1371/journal.pcbi.0020013

Abbreviations: ADMR, adrenomedullin receptor; CM, comparative modeling; GPCR, G protein–coupled receptor; NMR, nuclear magnetic resonance; PDB, Protein Data Bank; RH, rhodopsin; TASSER, threading assembly refinement; TM, transmembrane

* To whom correspondence should be addressed. E-mail: skolnick@gatech.edu

Synopsis

G protein–coupled receptors (GPCRs) are a large superfamily of integral membrane proteins that transduce signals across the cell membrane. Because of the breadth and importance of the physiological roles undertaken by the GPCR family, many of its members are important pharmacological targets. Although the knowledge of a protein's native structure can provide important insight into understanding its function and for the design of new drugs, the experimental determination of the three-dimensional structure of GPCR membrane proteins has proved to be very difficult. This is demonstrated by the fact that there is only one solved GPCR structure (from bovine rhodopsin) deposited in the Protein Data Bank library. In contrast, there are no human GPCR structures in the Protein Data Bank. To address the need for the tertiary structures of human GPCRs, using just sequence information, the authors use a newly developed threading-assembly-refinement method to generate models for all 907 registered GPCRs in the human genome. About 820 GPCRs are anticipated to have correct topology and transmembrane helix arrangement. A subset of the resulting models is validated by comparison with mutagenesis experimental data, and consistent agreement is demonstrated.

ing (comparative modeling [CM]) [11,12], threading [13,14], and ab initio folding [15–17]—CM, which builds models by aligning the target sequence to an evolutionarily related template structure, provides the most accurate models. However, its success is largely dictated by the evolutionary relationship between target and template proteins. For example, for proteins with greater than 50% sequence identity to their templates, CM models tend to be quite close to the native structure, with a 1-Å root-mean-squared-deviation (RMSD) from native for their backbone atoms, comparable to low-resolution X-ray and NMR experiments [12,18]. When the sequence identity drops below 30%, termed the "twilight zone," CM model accuracy sharply decreases because of the lack of a significant structure match and substantial alignment errors. Here, the models provided by CM are often closer to the template on which the model is based rather than the native structure of the sequence of interest. This has been a significant unsolved problem [19]. Among all registered human GPCRs, there are only four sequences that have a sequence identity to bovine RH greater than 30%. Ninety-nine percent of human GPCRs, with an average sequence identity to bovine RH of 19.5%, lie outside the traditional comparative modeling regimen [9].

Recently [14,17,20,21], we developed the threading assembly refinement (TASSER) methodology, which combines threading and ab initio algorithms to span the homologous to nonhomologous regimens. In a large-scale, comprehensive benchmark test of 2,234 representative proteins from the Protein Data Bank (PDB) [22], after excluding templates having greater than 30% sequence identity to the target, two thirds of single domain proteins can be folded to models with a C_α RMSD to native of less than 6.5 Å [20,21]. As a significant advance over traditional homology modeling, many models (including membrane proteins) are improved with respect to their threading templates (858 of 2,234 targets have an RMSD improvement of greater than 1.5 Å).

In the absence of additional GPCR crystal structures, computer-based modeling may provide the best alternative to obtaining structural information [23–28]. In this work, we exploit TASSER to predict tertiary structures for all 907 GPCR sequences in the human genome that are less than 500 amino acids in length. Only the sequence of the given GPCR is passed to TASSER and no other extrinsic knowledge (e.g., active sites and binding regions, experimental restraints, etc.) is incorporated into our structure prediction approach. Because the rearrangements of TM helices from RH may occur for nonhomologous GPCRs, the ability to refine templates is the most important advantage of using TASSER in comprehensive GPCR modeling. Also, distinct from many other GPCR modeling methods that only attempt to model the TM helical regions [27,29,30], TASSER generates reasonable predictions for the loop regions. In benchmark tests [21], for 39% of loops of four or more residues, TASSER models have a global RMSD less than 3 Å from native. In contrast, using the widely used homology modeling tool, MODELLER [11,12], the percentage of loops with this accuracy is 12% [20]. If one considers only the accuracy of the loop conformation itself (and neglects its orientation relative to the remainder of the protein), then 89% of the TASSER-generated loops have a local RMSD of less than 3 Å, and the average RMSD for loops up to 50 residues is below 4 Å. This is especially important in GPCR modeling as the extracellular loops are often critical in determining ligand specificity [31–33]. Therefore, full-length TASSER models offer substantial advantages over traditional comparative modeling methods and are likely to be of greater aid in understanding the ligand and signaling interactions of GPCRs.

Results

Application of TASSER to Membrane Proteins

Two forms of TASSER were developed for this study that slightly differ from our previously published work [14,17,20,21]. The first form of TASSER was extended to explicitly model TM proteins by including a "hydrophilic inside" potential for predicted TM regions as described in Materials and Methods. Modeling bovine RH with this form of TASSER demonstrates the effectiveness of this approach. On excluding homologous structures whose sequence identity greater than 30%, PROSPECTOR_3 identified three templates, 1pv6B (lactose permease), 1b3uA (protein phosphatase 2A), and 1a8hA (methionyl-tRNA synthetase), with Z-scores of 8.1, 8.7, and 5.3, respectively; bovine RH is therefore assigned as a medium/hard target. After the TASSER simulation, 76% of the structures from the 14 lowest temperature replicas are found inside a cluster with an RMSD cutoff of 8 Å. The average RMSD of these structures to the cluster centroid is 4.2 Å, which gives a C-score of 0.45. Of targets with this score, 82% are foldable according to the PDB benchmark [20]. In Figure 1, we show the comparison of both threading templates and the model of highest structure density with respect to the crystal structure. An RMSD of 4.6 Å from native for the final model is obtained if we superimpose all 338 C_α atoms (ten residues are absent in the crystal structure). The major modeling errors are in the N- and C-termini and the C3 loop. If we excise the tails and superimpose the model onto the core region (residues 32 to 323) of the native structure, the RMSD between the model and native structure is 3.3 Å. When we consider only the TM helix region, that is, TM1 (35 to 64), TM2 (71 to 100), TM3 (107 to 139), TM4 (151

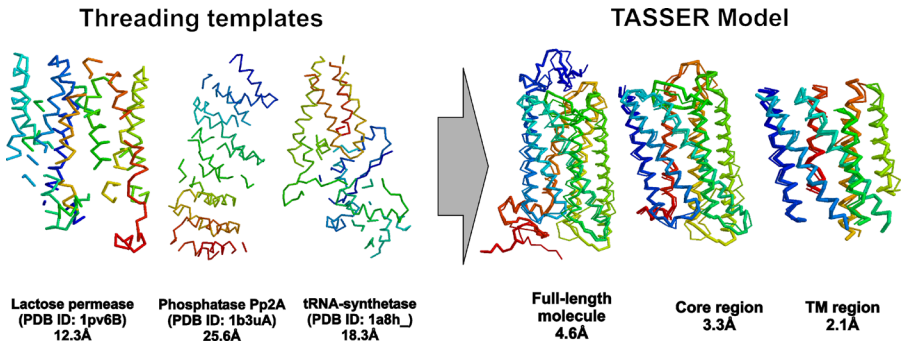

Figure 1. Initial Templates from PROSPECTOR_3 and the Final TASSER Model of Highest Cluster Density Superposed on the Bovine RH Crystal Structure
Blue to red runs from N- to C-terminus. The numbers are the RMSD to native. Images are from RASMOL [120].
DOI: 10.1371/journal.pcbi.0020013.g001

to 173), TM5 (200 to 225), TM6 (247 to 277), and TM7 (286 to 306), the RMSD is 2.1 Å.

A second integrated form of TASSER was constructed that incorporates a TM potential but selectively applies it without prior knowledge as to whether a target sequence is a membrane protein. Application of this integrated potential to a benchmark set of 38 membrane proteins (excluding all templates with greater than 30% identity in the aligned region) results in 17 targets with an RMSD to native less than 6.5 Å and an average improvement over the template alignment of 4.9 Å with 97% of targets showing an improvement compared to the starting template (Figure 2). A detailed list of the threading templates and final model information for the 38 membrane proteins is presented at Table 1. It should also be noted that more than 60% of the structures in the benchmark were proteins crystallized as part of large heteroprotein complexes. Applying this to the four other known seven-TM proteins in the PDB database, archeorhodopsin (1uaz), sensory rhodopsin (1jgj), halorhodopsin (1e12), and bacteriorhodopsin (1ap9), yields final models with RMSDs to native of 2.66, 1.25, 2.39, and 1.86 Å, respectively (Table 1). In all cases, TASSER refined the starting template

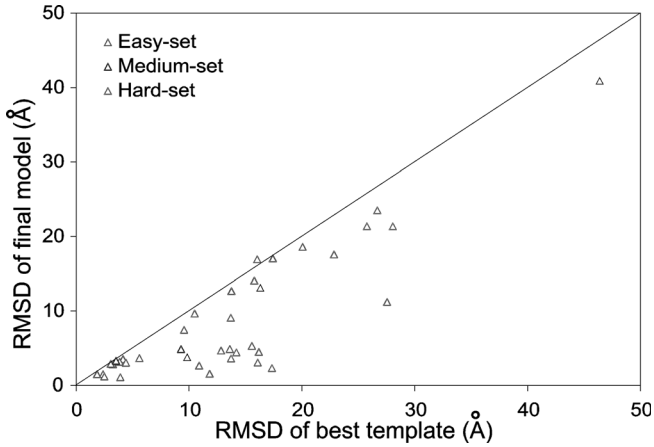

Figure 2. Application of TASSER to Membrane Proteins
TASSER was applied to a benchmark set of 38 membrane proteins with structures in the PDB. RMSD to native for final models of TASSER versus RMSD to native for initial templates from PROSPECTOR_3. All points beneath the 45° line indicate an improvement in the TASSER model over the initial template. All template alignments with a sequence identity greater than 30% were excluded from consideration.
DOI: 10.1371/journal.pcbi.0020013.g002

closer to native, with archeorhodopsin showing a change in RMSD from 19.78 Å for the highest scoring template to 1.22 Å over the same aligned region in the final model, sensory rhodopsin showing an improvement in RMSD from 2.50 Å to 1.18 Å, halorhodopsin showing an improvement in RMSD from 1.84 Å to 1.48 Å, and bacteriorhodopsin showing an improvement in RMSD from 2.37 Å to 1.49 Å.

As indicated in Table 1, there is unfortunately no clear pattern with regard to the type of proteins where TASSER modeling will succeed, because its successes and failures are scattered among the different types of membrane proteins (including α- and β-proteins). In fact, there are two factors contributing to the success of TASSER modeling. First, the dominant factor is the correct identification of analog templates from the threading algorithm [14]. Reasonable threading alignments provide a good starting point and framework for the follow-up TASSER refinement. Second, the composite and optimized knowledge-based TASSER force field contributes to the refinement of the models. The result of the final predictions is a combination of complex threading and simulation procedures, which prohibits the induction of a simple and explicit rule for when TASSER will succeed. Nevertheless, most proteins with a TM helical topology were well modeled by TASSER, a feature that is important for GPCR modeling. This may be due to the well-constructed sequence profiles from the extensive set of helical proteins in the sequence database, because PROSPECTOR__3 partly relies on a profile-profile alignment and the TASSER potential uses the short-range correlations identified by sequence profile matches.

One of the difficulties in validating GPCR models is the paucity of experimental evidence that would provide a strong validation or invalidation of a given model. However, by providing a detailed benchmark of membrane proteins including seven-TM proteins and bovine RH itself, we have clearly demonstrated the ability of TASSER to refine membrane structures from low sequence–identity templates to structures that are closer to the native structure in an automated fashion. The automated nature of this approach offers a potential advantage over many other human expert–based methods that may introduce biases by a priori assuming specific structural characteristics or restraints.

Sequence Clustering of Human GPCRs

Sequence analysis estimates that there are about 950 GPCRs in the human genome [3]. Combining the registered

Table 1. Modeling Result of the Benchmark Set of PDB Membrane Proteins

Target	Class	Template[a]	R_T (Å)[b]	$R_{M\text{-}ali}$(Å)[c]	$R_{M\text{-}all}$(Å)[d]	C-Score
1a87_	Easy	1cii_	46.38	40.91	40.35	−0.7
1aigL	Easy	1izlA	15.56	5.30	6.03	1.5
1aigM	Medium	1izlA	9.82	3.79	10.29	1.1
1ap9_	Easy	1jgjA	2.37	1.49	1.86	2.1
1bccF	Hard	1ps6A	16.30	13.12	13.37	−1.1
1bccH	Easy	1ga3A	9.55	7.46	8.68	1.0
1bh3_	Medium	2por_	9.28	4.87	4.97	1.2
1bl8A	Easy	2a79B	4.14	3.52	3.62	2.6
1bxwA	Hard	1p4tA	16.18	4.50	4.61	0.2
1e12A	Easy	1jgjA	1.84	1.48	2.38	2.1
1ezvH	Hard	1zpyA	17.34	2.32	2.29	0.9
1fftC	Easy	1occC	3.25	2.83	2.83	3.5
1fqyA	Easy	1fx8A	4.41	3.04	3.09	1.7
1fx8A	Easy	1fqyA	3.84	3.17	4.08	2.2
1gu8A	Easy	1e12A	3.90	1.09	7.17	2.1
1i78A	Hard	1k24A	28.05	21.37	25.04	−3.5
1jb0C	Medium	1clf_	3.53	3.31	8.43	1.0
1jb0J	Easy	1ug2A	13.61	4.88	4.96	1.2
1k24A	Hard	1i78A	26.68	23.52	23.15	−1.8
1kqfC	Hard	1cd5A	21.05	17.65	19.56	−1.8
1kzuA	Hard	1ahl_	14.19	4.43	5.36	0.4
1kzuB	Hard	1akhA	11.81	1.56	2.87	0.7
1lghB	Easy	1ocp_	13.75	12.67	13.51	−1.2
1lkfA	Easy	7ahlA	17.41	17.04	16.89	−3.2
1occC	Easy	1fftC	3.06	2.90	17.41	−0.8
1occE	Hard	1ijxA	13.70	9.09	10.38	−0.9
1occH	Hard	1b0yA	10.90	2.66	8.16	0.2
1occJ	Hard	1tig_	15.78	14.07	14.60	−0.8
1occL	Hard	1nz9A	10.50	9.65	13.07	−0.1
1occM	Hard	1cjgA	13.72	3.61	3.60	0.3
1orqC	Easy	2a79B	20.05	18.60	18.62	−1.5
1qcrK	Hard	1b0xA	12.83	4.68	5.05	0.1
1qd5A	Hard	1x8mA	27.54	11.22	11.28	−1.7
1qj8A	Easy	1p4tA	5.60	3.67	3.87	2.2
1qlaC	Hard	1ut9A	25.75	21.38	23.02	−1.6
1qleD	Hard	1t06A	16.08	3.06	4.044	0.3
1uazA	Medium	1syyA	19.78	1.22	2.659	1.0
7ahlA	Easy	1lkfA	16.03	16.92	19.55	−3.1

[a]The PDB ID of the best template with the lowest RMSD to native. All templates with sequence identity to the target greater than 30% were excluded.
[b]RMSD of the best template.
[c]RMSD of the TASSER models calculated in the threading aligned region.
[d]RMSD of the TASSER models calculated for the whole chain.
DOI: 10.1371/journal.pcbi.0020013.t001

entries in the http://www.gpcr.org/7tm/htmls/entries.html and http://www.expasy.org/cgi-bin/lists?7tmrlist.txt databases (February 2004 release), we find a total of 907 human GPCR sequences less than 500 residues in length. To establish their evolutionary distance, we made an all-against-all sequence comparison and grouped them into clusters based on their sequence identities. Four hundred forty-six GPCRs belong to the same sequence cluster with greater than 30% sequence identity; 384 of these are olfactory receptors, the largest subfamily in class A GPCRs [1,2]. The second largest cluster has 38 GPCR sequences, of which half are chemokine receptors. Three hundred sixty-five GPCRs belong to 68 smaller clusters with two to 30 members, including the four-member cluster homologous to bovine RH. The remaining 58 GPCRs are orphans with no partners having sequence identity greater than 30%. If we use sequence cutoffs of 20%, 25%, 35%, and 40%, there are 664, 477, 377, and 308 members in the largest sequence cluster, respectively. These data demonstrate the high sequence (and therefore structure) diversity among the GPCRs. If the assumption is made that GPCRs should all contain seven-TM regions—which may be incorrect—better alignments should be constructed by identifying helical regions explicitly. However, these sequence diversity data strongly suggest that direct comparative modeling with the bovine RH structure alone is highly unlikely to capture the nature of the structural differences among GPCRs not only in the highly diverse loop regions but within the core TM regions, too.

Threading Results

On threading the 907 GPCR sequences through our template library, a representative protein set covering PDB at the level of 35% sequence identity, PROSPECTOR_3 [14] assigns 778 sequences as easy targets, with average alignment coverage of 78%. This fraction of easy target assignment (about 86%) is significantly higher than in the PDB benchmark (about 67%) [20,21] and partly reflects the ability of PROSPECTOR_3 to detect the seven-TM helix bundle fold.

One hundred twenty-nine sequences are assigned as medium/hard targets with average alignment coverage of 67%.

The average sequence identities between the target and template are 17.8% and 15.5% for the easy and medium/hard targets, respectively. Despite these low sequence identities, there is some correlation between easy/medium/hard assignments and the size of sequence clusters that partially reflects the sensitivity of the sequence profile term in PROSPECTOR_3. Among the 48 sequence clusters with three or more members, all members in 40 of the clusters are easy targets. The 129 medium/hard targets are populated in a few sequence clusters: 84 of 129 of the medium/hard targets are olfactory receptors in sequence cluster I, and 17 of 129 are orphan GPCRs.

For most easy targets (652/778), PROSPECTOR_3 hits at least one of four seven-TM proteins (Sensory Rhodopsin, Halorhodopsin, Bacteriorhodopsin, or Bovine Rhodopsin) as templates. Although further refinement of the core region and the ab initio prediction of the loop conformations are needed, these alignments provide a reasonable initial conformation for TASSER. In fact, even for proteins that do not hit these four templates, due to TASSER refinement, many are predicted to have the TM helix topology through a fully automated procedure. As shown below, there are 862 cases where the GPCR model has a typical TM helix topology but only 744 targets have these four TM helical proteins as starting templates.

Confidence Score of Predicted Models

In Figure 3, the distribution of C-score, defined in Equation 1, for all 907 GPCRs is shown along with the corresponding C-score histogram of the PDB benchmark proteins [20]. Since the quality of threading templates is better for the GPCR proteins (reflected by the larger fraction of Easy targets and higher alignment coverage), many more GPCR models are populated at high C-scores.

In the PDB benchmark [20,21], for both globular and membrane proteins, there is a strong correlation between the C-score and the success rate of TASSER. For the 2,234 benchmark proteins ranging in size from 41 to 300 residues, the correlation coefficient between C-score and RMSD of the first model (corresponding to the most populated cluster) to native is 0.73. Of 38 membrane proteins in the benchmark, the correlation coefficient is 0.74, indicating that the TASSER confidence scoring system is directly applicable to TM proteins. The data in Figure 3 therefore imply that 819 of our GPCR models should have the correct topology. That is, for about 90% of the cases, at least one model in the top five predictions should have a core-region with an RMSD below 6.5 Å. There are 782, 749, and 698 cases with C-scores above 0.5, 1.0, and 1.5, respectively; in the benchmark, these C-scores correspond to a TASSER success rate of 94%, 97%, and 98%, respectively. Here, we note that a low RMSD just indicates the correctness of the overall topology of the helical arrangements. But the details of the loop regions and especially the ligand-binding sites may still be inaccurate. Further refinement at an atomic level as well as including the binding ligands in the modeling may be helpful.

It should be mentioned that although TASSER generates high confidence models for a substantial amount of medium/hard targets, the majority of the high C-score models are from easy targets. For example, among 749 targets with C-

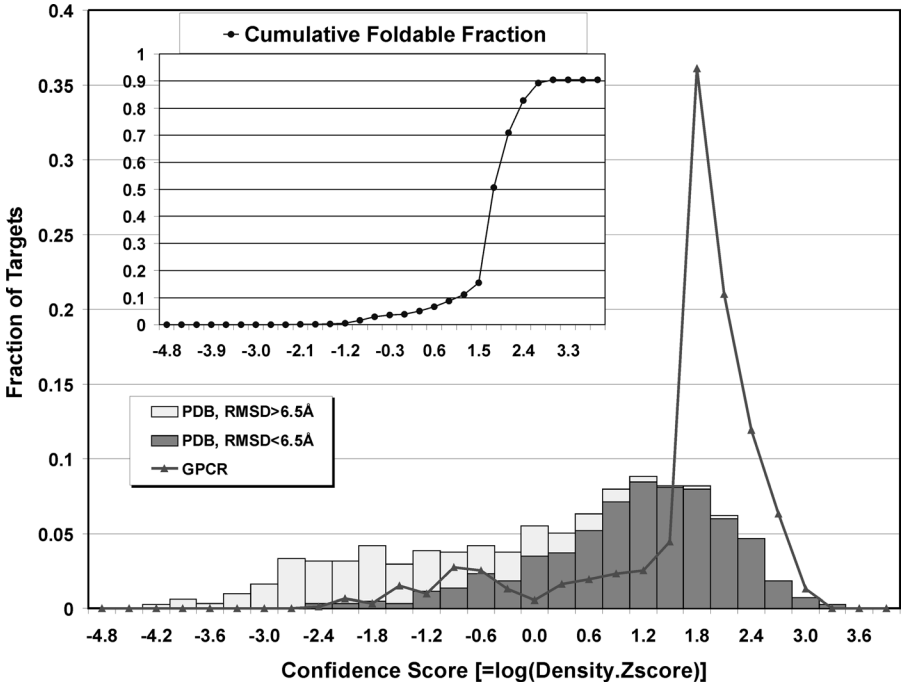

Figure 3. C-Score Distribution of the Predicted Models for the 907 GPCR Sequences

The C-score histogram for the PDB benchmark proteins [20] is shown for comparison, where dark gray denotes those models with an RMSD less than 6.5 Å to native and the light gray those models whose RMSD greater than 6.5 Å. The C-score is defined as in Equation 1. Inset: The cumulative foldable fraction calculated under the assumption that the GPCR proteins have the same correlation between success and C-score as that of the PDB benchmark proteins.

DOI: 10.1371/journal.pcbi.0020013.g003

score greater than 1.0, 659 are from easy targets. This correlation indicates that although TASSER has the ability of structural refinement, the overall success still strongly relies on the quality of threading templates [34]. Furthermore, models generated with the explicit membrane potential showed little difference from those generated with the integrated form of TASSER (average TM score, 0.91; average RMSD, 1.8 Å).

Conformational Changes from the Bovine RH Template

One of the major differences of the current approach from traditional CM methods is that TASSER refines the topology of threading alignments by rearranging the continuous fragments, while CM builds the model through optimally satisfying the restraints of the template structures. This results in the best CM models having the smallest variations from their initial template. Given the low sequence identity among GPCRs as a big family, one might expect significant differences from bovine RH, the only template available for CM methods. Thus, an interesting question is the extent to which TASSER has changed the conformation with respect to the initial template. In Figure 4A, we take all targets where TASSER employed bovine RH as an initial template and when the final model has a C-score greater than 1.3 and calculate the average distances of the residues of the final model from the corresponding residues in the bovine RH template according to the PROSPECTOR_3 alignment. On average, most residues in the TASSER model are greater than 4 Å away from the threading alignments with the maximum conformational changes in the loop and tail regions. In Figure 4B, we also show the helix angle changes of the predicted models with respect to bovine RH after superposition with TM-align [35]. Obviously, these conformational changes are significantly larger than the inherent resolution of TASSER modeling—as shown in the green triangles in Figure 4A and 4B; if we model bovine RH using its own crystal structure as the template, the overall RMSD of the model is 0.49 Å, with the observed variation along the RH template significantly smaller than the predicted average displacement for the other GPCR proteins. This degree of conformational change from the template is higher than could be expected by using a comparative modeling approach. Based on our previous benchmark and blind test results [20,21,36,37], most of the conformational deviations from the templates are in the correct direction toward native structures. For example, when starting from threading templates with a 4 to 5 Å RMSD to native, 58% of targets improve by at least 1 Å; when starting from a good threading template with a 2 to 3 Å RMSD to native, 43% of targets have at least a 0.5 Å improvement [20]. Even starting from the best structure alignments, similar improvements of final models relative to templates have been demonstrated (e.g., starting from initial structure alignments with an RMSD of 2 to 3 Å, 61% of targets have at least a 0.5 Å improvement) [36]. These data give us confidence that the observed deviations from the bovine RH template are most likely in the direction toward their native state.

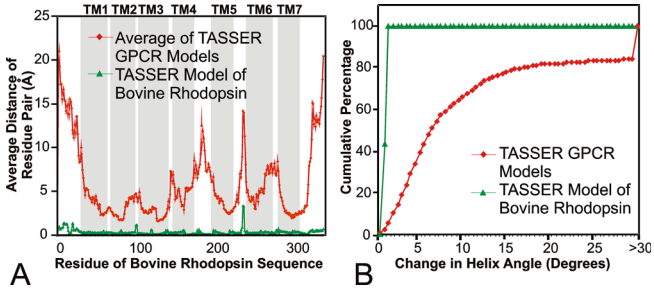

Figure 4. Conformational Changes of the Predicted TASSER Models from the Crystal Structure of Bovine RH

Data are the average from those targets where bovine RH is a template with C-score greater than 1.3 (red diamonds). The green triangles denote the TASSER model for bovine RH when bovine RH itself is used as the template (ten missed residues in 1f88 are inserted in the TASSER modeling). This shows the inherent resolution of the TASSER model.
(A) Average distance of each residue of the TASSER models from the bovine RH template along the sequence. TM helices are marked in gray.
(B) Percentage of all helices with helix angle changes below the indicated thresholds.
DOI: 10.1371/journal.pcbi.0020013.g004

Number of TM Helices

Using an automatic procedure to identify TM helices by structurally aligning the models to a long helix, we can count the number of TM helices in the predicted models. Consistent with the cell membrane thickness, these are typically 17 to 25 residues long [38]. Among the 907 GPCRs, 740 have the seven-TM helix bundle topology. Ten GPCRs have eight long helices, where, as visually confirmed, the eighth helix is located in a tail outside the seven-TM helix bundle. We also checked by visual inspection all other 157 targets that have fewer than seven helices. Most have shorter sequence lengths and a regular TM-like topology. In general, these are truncated fragments of complete GPCR sequences [39,40].

There are 45 sequences whose global topology is not TM helix–like. Most have zero- to three-long, non-TM helices. Sixteen of these are incomplete or alternately spliced transcripts; most are missing the majority of their TM regions; three (Q8TDU0, Q8TDV3, Q96HT6) do not appear to be GPCRs at all based on sequence analysis [41]; two (Q9HC23 and P06850) are ligands misannotated as GPCRs [42,43]. The remainder may represent an incorrect TASSER prediction, since TASSER does not have a trustable C-score for many of these targets. In fact, only four of these targets have a C-score greater than 1, including a target misannotated as a GPCR (Q9HC23) and three sequences (Q16503, Q96HT6, and Q99997) that are fragments of N-terminal domains and do not include the RH portions of the target sequence [32].

Structural Clusters of GPCR Models

Although there is little experimental evidence with which to directly test the validity of TASSER GPCR models, there exist indirect means of increasing the confidence in our predictions. First, an extensive membrane benchmark set from the PDB can be used to verify that TASSER can perform accurately on similar proteins. Second, we can check the self-consistency of the models under the assumption that closely related GPCRs or those with similar ligand specificities should in general adopt structures that are most similar to one another. To examine this, we first applied TM-align [35] to perform all-against-all structural alignments for the core regions of the predicted GPCR models and clustered the models based on their structural similarity. The average pairwise TM-score (a measure of fold consistency that ranges

from 0 to 1, with 1 being identical and below 0.17 being random [44] and that should not be confused with TM regions) for all 907 targets is 0.71, with an average RMSD of 3.1 Å with over 82% alignment coverage. These data demonstrate the strong structural conservation of the characteristic seven-TM helix topology. The conformational variance arises mainly from differences in TM helix packing and local helix kinks (Figure 4). Using a high TM score cutoff of 0.95, 228 GPCRs are clustered into 35 clusters; all other GPCRs have no structural analogs at this high TM score cutoff.

In Table 2, we present the top ten cluster results, ranked by the number of cluster members. There is a very strong tendency for GPCR function conservation within a given structure cluster. For example, all 59 members in the first structural cluster are in the olfactory II family according to their Swiss-Prot assignments [39,40]. There is no olfactory GPCR in the second cluster; but all 51 members belong to class A (or putative class A) RH-like GPCRs. In the third cluster, all ten members are chemokine receptors, a subfamily of peptide receptors. This demonstrates a consistency of structures with similar function.

Interestingly, the degree of sequence conservation varies among the structure clusters. For example, in cluster 7 where all members are Mas or Mas-related receptors, the average pairwise sequence identity is 87%. In contrast, in cluster 2, the average sequence identity is 23%, much lower than the permissive threshold allowed for robust sequence-based function inference [45]. In cluster 2, the lowest pairwise sequence identity, between P04001-P43116, is 13%, but the models for these two GPCRs have a TM-score of 0.95 and an RMSD of 1.2 Å over 97.4% of the residues, consistent with the observation that structure is more conserved than sequence in evolution [46]. These examples of sequence divergence with structure convergence also appear in other clusters (Table 2). It seems suggestive that the global structural information in the GPCR models may be a useful complement to sequence-based functional analysis [6].

As an additional means of examining the consistency of the TASSER models, we examined whether the GPCR subfamily could be determined based on structure alone. A benchmark set of GPCR models with a C-score greater than 1.3 that were part of GPCR subfamilies with similar or identical ligand specificities was constructed including adenosine, angiotensin, chemokine, endothelial cell differentiation, galanin, melanocortin, opioid, P2Y nucleotide, prostaglandin, somatostatin, trace amine, and arginine vasopressin subfamilies. In total, the set consisted of 72 receptors and 12 subfamilies. N- and C-terminal tails were removed since TASSER tends to model these regions poorly. Each structure in the set was compared by TM-align to each other structure in the set. In 75% of the cases, the structure with the highest TM-score belongs to the correct subfamily (86% of cases have a correct subfamily member with the four highest scoring structures). While standard phylogenetic analysis of the peptide sequences alone would yield correct results (the average sequence identity between any structure and other members of its subfamily is 40.5%), this result does indicate a high degree of consistency amongst the TASSER model structures.

Consistency of Models with Mutagenesis Studies

Although no solved X-ray or NMR structure is available for human GPCR sequences, numerous affinity labeling and site-directed mutagenesis experiments have been performed to identify critical residues and motifs that participate in ligand binding [2,31,47]. These data provide useful clues about the spatial contacts of the active site residues by which we can check to see if our models are consistent. We compared all TASSER models with C-scores greater than 1.3 to available mutagenesis studies including complement 5a receptor, thyroid-releasing hormone receptor, angiotensin receptor 1, adenosine 3 receptor, chemokine receptors, opioid receptors, formyl peptide receptor, thromboxane A_2 receptor, neuromedin B receptor, melatonin 2 receptor, gonadotropin-releasing hormone receptor, and neuropeptide Y receptors [48–111]. A subset of these receptors is shown in Figure 5 with critical residues marked. Excluding N- and C-terminal tails, we have not found any data that would invalidate our TASSER models.

Prediction of Ligand Specificity for an Orphan Receptor

While the ligand binding affinities of GPCRs tend to closely follow the sequence-based phylogenetic placement, there are

Table 2. Top Ten Structural Clusters of GPCR Models

Cluster	N[a]	<id>[b]	id_m[c]	ID[d]	TM/RMSD/Cov[e]	Function[f]
1	59	0.38	0.23	Q9H342-Q8NG82	0.96/1.1 Å/0.995	Olfactory II fam 2–8
2	51	0.23	0.13	P04001-P43116	0.95/1.2 Å/0.974	Class A nonolfactory
3	10	0.33	0.25	P46092-Q16570	0.97/0.9 Å/0.986	Chemokine
4	10	0.34	0.26	Q8NGI0-Q8NGK9	0.96/1.2 Å/0.993	Olfactory
5	8	0.46	0.27	O95499-Q9Y5P1	0.95/1.4 Å/0.988	Olfactory
6	6	0.38	0.30	Q8NGI1-Q8NH67	0.97/0.9 Å/0.990	Olfactory
7	5	0.87	0.79	Q8TDD6-Q96LB2	0.99/0.7 Å/0.998	Mas and Mas related
8	5	0.50	0.41	Q8NGP3-Q8NGV6	0.97/1.2 Å/0.998	Olfactory
9	5	0.45	0.39	Q8NGK4-Q9Y5P0	0.97/1.0 Å/0.997	Olfactory
10	4	0.49	0.40	Q8NGD5-Q8NGL9	0.97/1.1 Å/0.996	Olfactory

[a]Number of members in the clusters.
[b]Average pairwise sequence identity of the GPCRs in the clusters.
[c]The minimum pairwise sequence identity of the GPCRs in the clusters.
[d]SWISS-PROT ID of the GPCRs that have the minimum sequence identity as in column 3.
[e]TM-score, RMSD, and the structural alignment coverage by TM-align for the predicted models of the GPCR pairs at column 4.
[f]Functional descriptions of the GPCR members in the cluster, from the annotation in SWISS-PROT database [40].
DOI: 10.1371/journal.pcbi.0020013.t002

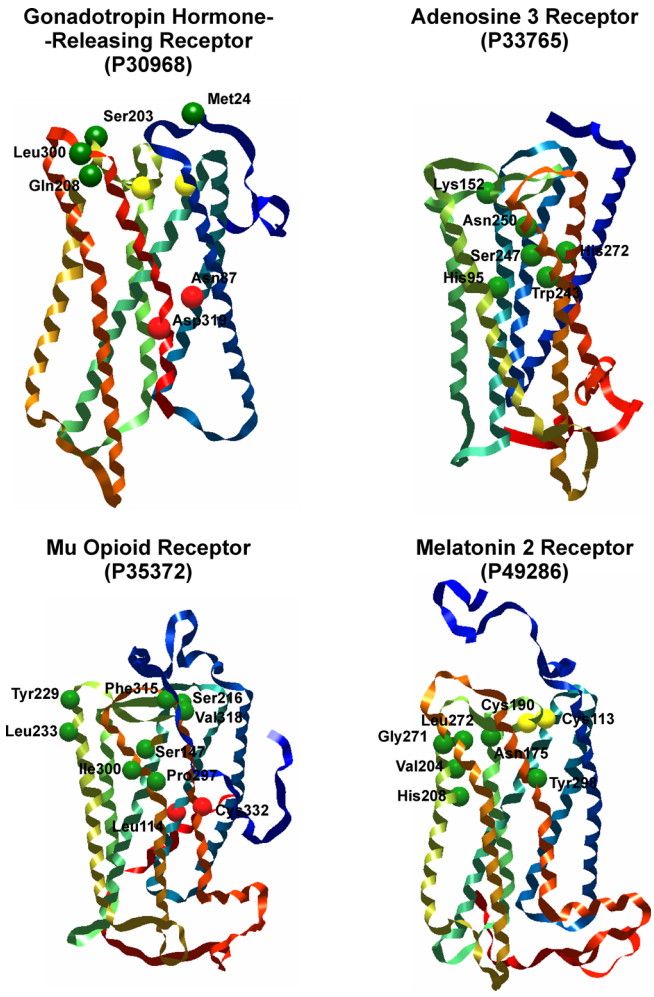

Figure 5. Consistency of Mutagenesis Studies with TASSER Predictions

TASSER models for gonadotropin hormone-release receptor, adenosine 3 receptor, mu opioid receptor, and melatonin 2 Receptor are shown with experimental determined residue interactions highlighted as spheres (green, ligand binding; yellow, disulfide bond; red, residue-residue interaction).

DOI: 10.1371/journal.pcbi.0020013.g005

instances where this is not the case. One such instance is the RDC1 receptor. After being initially identified, RDC1 remained an orphan receptor for 15 years. While RDC1 does not have a striking homology to any other GPCR, it appears to be most closely related to the adrenomedullin receptor (ADMR) and places consistently with it in phylogenetic studies [112,113]. However, RDC1 was recently shown to be a receptor for the chemokine CXCL12, which also binds the chemokine receptor CXCR4. Pairwise comparison of the TASSER model for RDC1 to all other 906 GPCR models without N- and C-terminal tails (which were generated prior to the discovery that RDC1 binds CXCL12) yields CXCR4 as the highest TM-score receptor, despite having a lower sequence identity through the same region. In fact, 63 other models have a higher TM-score to RDC1 than ADMR, many of which are other chemokine receptors, suggesting common structural characteristics that distinguish chemokine and adrenomedullin receptors. It is important to note that this is strictly based on a direct structural comparison with no explicit attention paid to residue identities. Not only does this provide evidence strengthening the validity of TASSER models, but it also suggests that these structures may also be applied toward resolving the ligand specificities of orphan receptors as well as toward classification of weakly homologous GPCRs in other species.

Discussion

By incorporating specific protein-membrane interactions into the TASSER force field, we have extended the TASSER threading-assembly-refinement methodology [20] and generated tertiary structure predictions for all 907 registered GPCRs in the human genome less than 500 amino acids long. Unlike traditional CM methods, TASSER does not require that the structures of homologous templates be solved, an essential attribute for the successful modeling of the whole set of human GPCR proteins, because most GPCRs have no close evolutionary relationship to proteins in the PDB.

Moreover, TASSER often refines the structures closer to native than the templates on which they are based [20,21]. This is particularly important for understanding the functional subtleties of the different classes of GPCRs when the models start from similar template alignments. These features have been demonstrated in the benchmark modeling of 38 representative medium-size membrane proteins from the PDB library, where TASSER has drawn the initial threading templates closer to native by an average RMSD of 4.9 Å in the threading aligned regions. An example of special interest is from RH of bovine rod outer segment membrane, the only solved GPCR protein. Excluding homologous proteins of sequence identity greater than 30% as well as bacteriorhodopsin, the threading program PROSPECTOR_3 [14] identifies three helical templates, all with global RMSD greater than 10 Å. After TASSER reassembly, the first model has a full-length RMSD to native of 4.6 Å, with the TM helix region having an RMSD of 2.1 Å. Recently, there have been many other attempts to model the tertiary structure of bovine RH. For example, Sale et al. [30] modeled the TM helix region using a statistical potential combined with 27 experimental distance constraints. They built a model with an RMSD of 3.2 Å to native in the TM region. Becker and coworkers [27,29] used PREDICT to model the TM region and generated a model whose TM region has an RMSD from native of 2.9 Å. Using MembStruk, Vaidehi et al. [114] built a model with an RMSD from native of 3.1 Å in the TM region and an RMSD from native of 8.3 Å for the full-length molecule. Compared with these results, the TASSER model is more accurate in the TM helix region, the loop/tail regions, and the full-length molecule.

Among the models generated for the 907 GPCR sequences, based on the confidence score established in comprehensive PDB benchmarking [20,21], 819 GPCRs are anticipated to have the correct global fold. Seven hundred fifty ORFs have the characteristic seven-TM helix topology, and 112 ORFs have the TM helix bundle topology with less than seven TM helices. There are 45 cases where TASSER generates non-TM helical models, primarily because these sequences come from periplasmic domain regions.

A quantitative structural comparison of the models from different GPCRs was performed by an all-against-all structural superposition. The average pairwise TM-score of the

907 GPCR models is 0.71, with an average 3.1 Å RMSD for 82% of residues in the core region. Using a restrictive TM-score cutoff of 0.95, the models tend to be grouped into structural clusters that have strong functional conservation, although sequences can be very divergent within the clusters. This is suggestive that structural information from the predicted GPCR models can be a useful complement to sequence-based functional analysis. It also demonstrates the robustness of TASSER, since structural convergence at low sequence identity is not built in but is a prediction.

A further validation of the predicted models includes structural consistency of GPCR subfamilies binding the same or similar ligands and consistency with mutagenesis studies. Furthermore, we demonstrate that the GPCR models can be more sensitive in determining ligand specificity than sequence-based methods, as is evidenced by the TASSER model of RDC1. Using sequence-based methods, RDC1 was expected to be an adrenomedullin receptor, since it shares its highest sequence identity and places phylogenetically with ADMR. However, RDC1 was recently shown to bind the chemokine CXCL12, whose only other known receptor is CXCR4. The TASSER model of RDC1 has as its closest structural neighbor, the model of CXCR4, further supporting the accuracy of TASSER models.

Comparative modeling approaches are useful in inferring the structures of sequences with greater than 30% sequence identity. They are also attractive because the computation resources required in generating these models are relatively small. However, 99% of GPCRs have a sequence identity less than 19.5% to the only solved GPCR structure, bovine RH. It is clear that comparative modeling alone would be unable to capture the range of structural diversity encompassed by the 907 receptors examined in this study. Alternatively, receptors can be modeled using specific restraints and assumptions that are assumed to be true for all GPCRs based on the solved rhodopsin structures at the risk of missing unforeseen structural characteristics of poorly characterized GPCRs. Threading using PROSPECTOR__3 along with TASSER has a demonstrated ability to construct accurate models of both membrane and globular proteins in a completely automated fashion with low-homology templates, thus providing an advantage over both the comparative modeling techniques and methods geared to strictly modeling GPCRs. While definitive validation of these structures is difficult given the paucity of clearly discriminating experimental evidence—the very reason why many have looked to predicted models in the first place—our benchmark studies of other membrane proteins and examination of the GPCR models for consistency with existing observations strongly suggests that these models are rather accurate.

The models presented here represent the most complete set of GPCRs models developed to date and offer a resource for ligand screening and other functional predications [27,115]. Given the extensive computational time required to generate these models (several decades, if run on a single processor), this study makes a resource available for experimental testing that would be infeasible for most experimental labs to generate independently. All predicted GPCR models, as well as follow-up functional analysis data, are available for noncommercial users on our Web site (http://www.bioinformatics.buffalo.edu/GPCR).

Materials and Methods

Template identification. For a given GPCR sequence, we run the threading program PROSPECTOR__3 to identify putative related template structures in the PDB. PROSPECTOR__3 is an iterative sequence/structure alignment approach [14]. On the basis of the score significance and the consensus of template alignments, proteins are categorized into easy, medium, and hard targets. These terms refer to the relative confidence in the accuracy of the predicted threading models. According to the benchmark, 80% of the threading-predicted alignments for the easy targets have an RMSD to native less than 6.5 Å in the aligned regions [20]. This alignment accuracy is essentially the same for both globular and membrane proteins [21]. For the medium/hard targets, the topology of the template is often correct, but the global alignment can be in error. Nevertheless, the local fragments from the template alignment can be utilized as building blocks in TASSER [20].

Substructure/fragment assembly. Continuous fragments (more than five residues) are excised from the five top scoring threading templates for the easy targets and up to the 20 top templates for the medium/hard targets. For the GPCR sequences, these fragments are mainly long TM helices that will be reassembled under the guide of the TASSER force field that consists of an optimized combination of a reduced knowledge-based potential [17] and consensus contact restraints from threading [14]. The loops connecting the helices are generated by the TASSER ab initio structure prediction procedure.

Conformational space is searched by the parallel hyperbolic Monte Carlo algorithm [116]. Depending on GPCR length, 40 to 80 replicas are used with larger molecules having more replicas. Two kinds of major conformational updates are implemented: Off-lattice movements of the template-excised fragments involve rigid translations and rotations controlled by the three Euler angles of each fragment. Lattice-confined loop residues are subject to two to six bond movements and multibond sequence shifts [17].

Extension of TASSER to membrane proteins. The hydrophobic interactions in the original TASSER force field are applied only to the loop/tail residues, which are assumed to be outside the membrane. For the putative TM helices, because of the hydrophobic membrane environment, a propensity for hydrophilic side chains to face to the interior of the protein is included. TM helices are assigned from PSIPRED [117]. We also tried other TM predictors, e.g., MEMSAT [38], but the differences are small. In general, the local geometry of a template-derived substructure remains similar to that in the template [20]. However, considering the variance of helix shape and the presence of local kinks along the TM helices, we allow a small bending deformation for the aligned TM helices. A strong penalty potential term of $E \sim \Delta RMSD^4$ is employed (the form of the fourth power is somewhat arbitrary but was chosen based on trial and error); $\Delta RMSD$ denotes the RMSD between the excised template substructure and the deformed substructure in the simulations.

Model selection and assessment. Trajectories of the 14 lowest temperature replicas are submitted to SPICKER [37] for structure clustering. SPICKER first identifies a center structure that has the most neighboring structures within an RMSD threshold R_{cut}. The first cluster is defined as a group of structures including the center structure and all its neighbors. The second cluster is similarly defined after all the structures in the first cluster have been removed, etc. R_{cut} is defined in an iterative way: The initial R_{cut} is set to 7.5 Å. If the structures are too tightly distributed, R_{cut} will gradually decrease until the first cluster contains less than 70% of the total number of structures or until R_{cut} is 3.5 Å. If the structures are too divergent, R_{cut} will gradually increase until the first cluster includes more than 15% of structures or until $R_{\mathrm{cut}} = 12$ Å.

To avoid distortions of clusters from disordered tail variations, a two-step clustering is implemented: We first run SPICKER on the structurally well defined core (the putative TM region based on PSIPRED) and the tail regions separately; then, we dock the conformations of the three separate parts (the two tails and the core) into the final full-length model based on the superposition of these regions onto the structure obtained by clustering the full-length conformations. Reduced models (C_αs and side-chain centers of mass) are provided from the clustered structures. Backbone and side-chain heavy atoms are added using PULCHRA [118].

The final models are ranked and assessed on the basis of the confidence score [20]:

$$\mathrm{C-score} = \ln\left(\frac{M}{\langle rmsd\rangle M_{\mathrm{tot}}}Z\right) \quad (1)$$

where M is the multiplicity of structures in a SPICKER cluster, M_{tot} is

the total number of structures submitted for clustering, and $\langle rmsd \rangle$ denotes the average RMSD of the structures relative to the cluster centroid. The logarithm in Equation 1 serves to expand the range of the C-score distribution of the predicted structures. If we define a correct fold as the one with an RMSD to native below 6.5 Å [119], the PDB benchmark results [20] show that for all targets with a C-score threshold of –0.5, the total false positive/false negative rate is 12.4%/14.7%.

Supporting Information

Accession Numbers

The Swiss-Prot (http://www.ebi.ac.uk/swissprot) accession numbers for the four sequences of registered human GPCRs that have a sequence identity to bovine RH greater than 30% are P03999, P04000, P04001, and P08100.

The Protein Data Bank (http://www.rcsb.org/pdb) accession number for the bovine RH crystal structure is 1f88A.

Acknowledgments

We thank Dr. A. Szilagyi for stimulating discussions and Drs. A. Arakaki and V. Grimm for help in preparing the GPCR sequences.

Author contributions. YZ, MED, and JS conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, and wrote the paper.

Funding. This research was supported in part by grants GM-37408 and GM-48835 of the Division of General Medical Sciences of the National Institutes of Health.

Competing interests. The authors have declared that no competing interests exist. ■

References

1. Watson S, Arkinstall S (1994) The G protein Linked Receptors Facts Book. New York: Academic Press. 427 p.
2. Flower DR (1999) Modelling G-protein-coupled receptors for drug design. Biochim Biophys Acta 1422: 207–234.
3. Takeda S, Kadowaki S, Haga T, Takaesu H, Mitaku S (2002) Identification of G protein-coupled receptor genes from the human genome sequence. FEBS Lett 520: 97–101.
4. Collins FS (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.
5. Drews J (2000) Drug discovery: A historical perspective. Science 287: 1960–1964.
6. Skolnick J, Fetrow JS, Kolinski A (2000) Structural genomics and its importance for gene function analysis. Nat Biotechnol 18: 283–287.
7. Kuhlbrandt W, Gouaux E (1999) Membrane proteins. Curr Opin Struct Biol 9: 445–447.
8. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, et al. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. Science 289: 739–745.
9. Archer E, Maigret B, Escrieut C, Pradayrol L, Fourmy D (2003) Rhodopsin crystal: New template yielding realistic models of G-protein-coupled receptors? Trends Pharmacol Sci 24: 36–40.
10. Moult J, Fidelis K, Zemla A, Hubbard T (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 53 (Suppl 6): 334–339.
11. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779–815.
12. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29: 291–325.
13. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253: 164–170.
14. Skolnick J, Kihara D, Zhang Y (2004) Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Protein 56: 502–518.
15. Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA (1999) Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci U S A 96: 5482–5485.
16. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268: 209–225.
17. Zhang Y, Kolinski A, Skolnick J (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophys J 85: 1145–1164.
18. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294: 93–96.
19. Tramontano A, Morea V (2003) Assessment of homology-based predictions in CASP5. Proteins 53 (Suppl 6): 352–368.
20. Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101: 7594–7599.
21. Zhang Y, Skolnick J (2004) Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophys J 87: 2647–2655.
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
23. Herzyk P, Hubbard RE (1995) Automated method for modeling seven-helix transmembrane receptors from experimental data. Biophys J 69: 2419–2442.
24. Filizola M, Perez JJ, Carteni-Farina M (1998) BUNDLE: A program for building the transmembrane domains of G-protein-coupled receptors. J Comput Aided Mol Des 12: 111–118.
25. Bissantz C, Logean A, Rognan D (2004) High-throughput modeling of human G-protein coupled receptors: Amino acid sequence alignment, three-dimensional model building, and receptor library screening. J Chem Inf Comput Sci 44: 1162–1176.
26. Bissantz C, Bernard P, Hibert M, Rognan D (2003) Protein-based virtual screening of chemical databases. II. Are homology models of G-protein coupled receptors suitable targets? Proteins 50: 5–25.
27. Shacham S, Marantz Y, Bar-Haim S, Kalid O, Warshaviak D, et al. (2004) PREDICT modeling and in-silico screening for G-protein coupled receptors. Proteins 57: 51–86.
28. Becker OM, Shacham S, Marantz Y, Noiman S (2003) Modeling the 3D structure of GPCRs: advances and application to drug discovery. Curr Opin Drug Disc Dev 6: 353–361.
29. Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A, et al. (2004) G protein-coupled receptors: In silico drug discovery in 3D. Proc Natl Acad Sci U S A 101: 11304–11309.
30. Sale K, Faulon JL, Gray GA, Schoeniger JS, Young MM (2004) Optimal bundling of transmembrane helices using sparse distance constraints. Prot Sci 13: 2613–2627.
31. Shi L, Javitch JA (2002) The binding site of aminergic G protein-coupled receptors: The transmembrane segments and second extracellular loop. Annu Rev Pharmacol Toxicol 42: 437–467.
32. Kunishima N, Shimada Y, Tsuji Y, Sato T, Yamamoto M, et al. (2000) Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. Nature 407: 971–977.
33. Du P, Salon JA, Tamm JA, Hou C, Cui W, et al. (1997) Modeling the G-protein-coupled neuropeptide Y Y1 receptor agonist and antagonist binding sites. Prot Eng 10: 109–117.
34. Zhang Y, Arakaki A, Skolnick J (2005) TASSER: An automated method for the prediction of protein tertiary structures in CASP6. Proteins: In press.
35. Zhang Y, Skolnick J (2005) TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33: 2302–2309.
36. Zhang Y, Skolnick J (2005) The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 102: 1029–1034.
37. Zhang Y, Skolnick J (2004) SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 25: 865–871.
38. Jones DT, Taylor WR, Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33: 3038–3049.
39. Horn F, Weare J, Beukers MW, Horsch S, Bairoch A, et al. (1998) GPCRDB: An information system for G protein-coupled receptors. Nucleic Acids Res 26: 275–279.
40. Bairoch A, Apweiler R (1998) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res 26: 38–42.
41. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, et al. (2005) CDD: A conserved domain database for protein classification. Nucleic Acids Res 33: D192–D196.
42. Pisarska M, Mulchahey JJ, Sheriff S, Geracioti TD, Kasckow JW (2001) Regulation of corticotropin-releasing hormone in vitro. Peptides 22: 705–712.
43. Chen J, Kuei C, Sutton S, Wilson S, Yu J, et al. (2005) Identification and pharmacological characterization of prokineticin 2beta as a selective ligand for prokineticin receptor 1. Mol Pharmacol 67: 2070–2076.
44. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57: 702–710.
45. Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333: 863–882.
46. Holm L, Sander C (1996) Mapping the protein universe. Science 273: 595–603.
47. Schwartz TW (1994) Locating ligand-binding sites in 7TM receptors by protein engineering. Curr Opin Biotechnol 5: 434–444.
48. Clement M, Martin SS, Beaulieu ME, Chamberland C, Lavigne P, et al. (2005) Determining the environment of the ligand binding pocket of the human angiotensin II type I (hAT1) receptor using the methionine proximity assay. J Biol Chem 280: 27121–27129.
49. Huang W, Osman R, Gershengorn MC (2005) Agonist-induced conforma-

tional changes in thyrotropin-releasing hormone receptor type I: Disulfide cross-linking and molecular modeling approaches. Biochemistry 44: 2419–2426.
50. Zhang Y, McCurdy CR, Metzger TG, Portoghese PS (2005) Specific cross-linking of Lys233 and Cys235 in the mu opioid receptor by a reporter affinity label. Biochemistry 44: 2271–2275.
51. Higginbottom A, Cain SA, Woodruff TM, Proctor LM, Madala PK, et al. (2005) Comparative agonist/antagonist responses in mutant human C5a receptors define the ligand binding site. J Biol Chem 280: 17831–17840.
52. Gavrilin MA, Gulina IV, Kawano T, Dragan S, Chakravarti L, et al. (2005) Site-directed mutagenesis of CCR2 identified amino acid residues in transmembrane helices 1, 2, and 7 important for MCP-1 binding and biological functions. Biochem Biophys Res Commun 327: 533–540.
53. Buck E, Bourne H, Wells JA (2005) Site-specific disulfide capture of agonist and antagonist peptides on the C5a receptor. J Biol Chem 280: 4009–4012.
54. de Mendonca FL, da Fonseca PC, Phillips RM, Saldanha JW, Williams TJ, et al. (2005) Site-directed mutagenesis of CC chemokine receptor 1 reveals the mechanism of action of UCB 35625, a small molecule chemokine receptor antagonist. J Biol Chem 280: 4808–4816.
55. Mazna P, Obsilova V, Jelinkova I, Balik A, Berka K, et al. (2004) Molecular modeling of human MT2 melatonin receptor: The role of Val204, Leu272 and Tyr298 in ligand binding. J Neurochem 91: 836–842.
56. Costanzi S, Mamedova L, Gao ZG, Jacobson KA (2004) Architecture of P2Y nucleotide receptors: Structural comparison based on sequence analysis, mutagenesis, and homology modeling. J Med Chem 47: 5393–5404.
57. Martin SS, Boucard AA, Clement M, Escher E, Leduc R, et al. (2004) Analysis of the third transmembrane domain of the human type 1 angiotensin II receptor by cysteine scanning mutagenesis. J Biol Chem 279: 51415–51423.
58. Reinhart GJ, Xie Q, Liu XJ, Zhu YF, Fan J, et al. (2004) Species selectivity of nonpeptide antagonists of the gonadotropin-releasing hormone receptor is determined by residues in extracellular loops II and III and the amino terminus. J Biol Chem 279: 34115–34122.
59. Tchilibon S, Kim SK, Gao ZG, Harris BA, Blaustein JB, et al. (2004) Exploring distal regions of the A3 adenosine receptor binding site: Sterically constrained N6-(2-phenylethyl)adenosine derivatives as potent ligands. Bioorg Med Chem 12: 2021–2034.
60. Geng L, Wu J, So SP, Huang G, Ruan KH (2004) Structural and functional characterization of the first intracellular loop of human thromboxane A2 receptor. Arch Biochem Biophys 423: 253–265.
61. Fadhil I, Schmidt R, Walpole C, Carpenter KA (2004) Exploring deltorphin II binding to the third extracellular loop of the delta-opioid receptor. J Biol Chem 279: 21069–21077.
62. Hernanz-Falcon P, Rodriguez-Frade JM, Serrano A, Juan D, del Sol A, et al. (2004) Identification of amino acid residues crucial for chemokine receptor dimerization. Nat Immunol 5: 216–223.
63. Berkhout TA, Blaney FE, Bridges AM, Cooper DG, Forbes IT, et al. (2003) CCR2: characterization of the antagonist binding site from a combined receptor modeling/mutagenesis approach. J Med Chem 46: 4070–4086.
64. Decaillot FM, Befort K, Filliol D, Yue S, Walker P, et al. (2003) Opioid receptor random mutagenesis reveals a mechanism for G protein-coupled receptor activation. Nat Struct Biol 10: 629–636.
65. Boucard AA, Roy M, Beaulieu ME, Lavigne P, Escher E, et al. (2003) Constitutive activation of the angiotensin II type 1 receptor alters the spatial proximity of transmembrane 7 to the ligand-binding pocket. J Biol Chem 278: 36628–36636.
66. Klco JM, Lassere TB, Baranski TJ (2003) C5a receptor oligomerization. I. Disulfide trapping reveals oligomers and potential contact surfaces in a G protein-coupled receptor. J Biol Chem 278: 35345–35353.
67. Gerdin MJ, Mseeh F, Dubocovich ML (2003) Mutagenesis studies of the human MT2 melatonin receptor. Biochem Pharmacol 66: 315–320.
68. Kokkola T, Foord SM, Watson MA, Vakkuri O, Laitinen JT (2003) Important amino acids for the function of the human MT1 melatonin receptor. Biochem Pharmacol 65: 1463–1471.
69. Gao ZG, Kim SK, Gross AS, Chen A, Blaustein JB, et al. (2003) Identification of essential residues involved in the allosteric modulation of the human A(3) adenosine receptor. Mol Pharmacol 63: 1021–1031.
70. Chaipatikul V, Loh HH, Law PY (2003) Ligand-selective activation of mu-oid receptor: demonstrated with deletion and single amino acid mutations of third intracellular loop domain. J Pharmacol Exp Ther 305: 909–918.
71. So SP, Wu J, Huang G, Huang A, Li D, et al. (2003) Identification of residues important for ligand binding of thromboxane A2 receptor in the second extracellular loop using the NMR experiment-guided mutagenesis approach. J Biol Chem 278: 10922–10927.
72. Guo W, Shi L, Javitch JA (2003) The fourth transmembrane segment forms the interface of the dopamine D2 receptor homodimer. J Biol Chem 278: 4385–4388.
73. Mseeh F, Gerdin MJ, Dubocovich MI (2002) Identification of cysteines involved in ligand binding to the human melatonin MT(2) receptor. Eur J Pharmacol 449: 29–38.
74. Hamdan FF, Ward SD, Siddiqui NA, Bloodworth LM, Wess J (2002) Use of an in situ disulfide cross-linking strategy to map proximities between amino acid residues in transmembrane domains I and VII of the M3 muscarinic acetylcholine receptor. Biochemistry 41: 7647–7658.
75. Auger GA, Pease JE, Shen X, Xanthou G, Barker MD (2002) Alanine scanning mutagenesis of CCR3 reveals that the three intracellular loops are essential for functional receptor expression. Eur J Immunol 32: 1052–1058.
76. Turek JW, Halmos T, Sullivan NL, Antonakis K, Le Breton GC (2002) Mapping of a ligand-binding site for the human thromboxane A2 receptor protein. J Biol Chem 277: 16791–106797.
77. Shapiro DA, Kristiansen K, Weiner DM, Kroeze WK, Roth BL (2002) Evidence for a model of agonist-induced activation of 5-hydroxytryptamine 2A serotonin receptors that involves the disruption of a strong ionic interaction between helices 3 and 6. J Biol Chem 277: 11441–11449.
78. Ward SD, Hamdan FF, Bloodworth LM, Wess J (2002) Conformational changes that occur during M3 muscarinic acetylcholine receptor activation probed by the use of an in situ disulfide cross-linking strategy. J Biol Chem 277: 2247–2257.
79. Kraft K, Olbrich H, Majoul I, Mack M, Proudfoot A, et al. (2001) Characterization of sequence determinants within the carboxyl-terminal domain of chemokine receptor CCR5 that regulate signaling and receptor internalization. J Biol Chem 276: 34408–34418.
80. Nardese V, Longhi R, Polo S, Sironi F, Arcelloni C, et al. (2001) Structural determinants of CCR5 recognition and HIV-1 blockade in RANTES. Nat Struct Biol 8: 611–615.
81. Conway S, Mowat ES, Drew JE, Barrett P, Delagrange P, et al. (2001) Serine residues 110 and 114 are required for agonist binding but not antagonist binding to the melatonin MT(1) receptor. Biochem Biophys Res Commun 282: 1229–1236.
82. Youn BS, Yu KY, Alkhatib G, Kwon BS (2001) The seventh transmembrane domain of cc chemokine receptor 5 is critical for MIP-1beta binding and receptor activation: role of MET 287. Biochem Biophys Res Commun 281: 627–633.
83. Katancik JA, Sharma A, de Nardin E (2000) Interleukin 8, neutrophil-activating peptide-2 and GRO-alpha bind to and elicit cell activation via specific and different amino acid residues of CXCR2. Cytokine 12: 1480–1488.
84. Tokita K, Hocart SJ, Katsuno T, Mantey SA, Coy DH, et al. (2001) Tyrosine 220 in the 5th transmembrane domain of the neuromedin B receptor is critical for the high selectivity of the peptoid antagonist PD168368. J Biol Chem 276: 495–504.
85. Chabot DJ, Broder CC (2000) Substitutions in a homologous region of extracellular loop 2 of CXCR4 and CCR5 alter coreceptor activities for HIV-1 membrane fusion and virus entry. J Biol Chem 275: 23774–23782.
86. Dragic T, Trkola A, Thompson DA, Cormier EG, Kajumo FA, et al. (2000) A binding pocket for a small molecule inhibitor of HIV-1 entry within the transmembrane helices of CCR5. Proc Natl Acad Sci U S A 97: 5639–5644.
87. Mirzadegan T, Diehl F, Ebi B, Bhakta S, Polsky I, et al. (2000) Identification of the binding site for a novel class of CCR2b chemokine receptor antagonists: Binding to a common chemokine receptor motif within the helical bundle. J Biol Chem 275: 25562–25571.
88. Zhou H, Tai HH (2000) Expression and functional characterization of mutant human CXCR4 in insect cells: role of cysteinyl and negatively charged residues in ligand binding. Arch Biochem Biophys 373: 211–217.
89. Han KH, Green SR, Tangirala RK, Tanaka S, Quehenberger O (1999) Role of the first extracellular loop in the functional activation of CCR2. The first extracellular loop contains distinct domains necessary for both agonist binding and transmembrane signaling. J Biol Chem 274: 32055–32062.
90. Miettinen HM, Gripentrog JM, Mason MM, Jesaitis AJ (1999) Identification of putative sites of interaction between the human formyl peptide receptor and G protein. J Biol Chem 274: 27934–27942.
91. Chabot DJ, Zhang PF, Quinnan GV, Broder CC (1999) Mutagenesis of CXCR4 identifies important domains for human immunodeficiency virus type 1 X4 isolate envelope-mediated membrane fusion and virus entry and reveals cryptic coreceptor activity for R5 isolates. J Virol 73: 6598–6609.
92. Zeng FY, Hopp A, Soldner A, Wess J (1999) Use of a disulfide cross-linking strategy to study muscarinic receptor structure and mechanisms of activation. J Biol Chem 274: 16629–16640.
93. Xu W, Ozdener F, Li JG, Chen C, de Riel JK, et al. (1999) Functional role of the spatial proximity of Asp114(2.50) in TMH 2 and Asn332(7.49) in TMH 7 of the mu opioid receptor. FEBS Lett 447: 318–324.
94. Suetomi K, Lu Z, Heck T, Wood TG, Prusak DJ, et al. (1999) Differential mechanisms of recognition and activation of interleukin-8 receptor subtypes. J Biol Chem 274: 11768–11772.
95. Sainz E, Akeson M, Mantey SA, Jensen RT, Battey JF (1998) Four amino acid residues are critical for high affinity binding of neuromedin B to the neuromedin B receptor. J Biol Chem 273: 15927–15932.
96. Wang ZX, Berson JF, Zhang TY, Cen YH, Sun Y, et al. (1998) CXCR4 sequences involved in coreceptor determination of human immunodeficiency virus type-1 tropism. Unmasking of activity with M-tropic Env glycoproteins. J Biol Chem 273: 15007–15015.
97. Mills JS, Miettinen HM, Barnidge D, Vlases MJ, Wimer-Mackin S, et al. (1998) Identification of a ligand binding site in the human neutrophil formyl peptide receptor using a site-specific fluorescent photoaffinity label and mass spectrometry. J Biol Chem 273: 10428–10435.
98. Rabut GE, Konner JA, Kajumo F, Moore JP, Dragic T (1998) Alanine substitutions of polar and nonpolar residues in the amino-terminal

domain of CCR5 differently impair entry of macrophage- and dualtropic isolates of human immunodeficiency virus type 1. J Virol 72: 3464–3468.
99. Dragic T, Trkola A, Lin SW, Nagashima KA, Kajumo F, et al. (1998) Amino-terminal substitutions in the CCR5 coreceptor impair gp120 binding and human immunodeficiency virus type 1 entry. J Virol 72: 279–285.
100. Miettinen HM, Mills JS, Gripentrog JM, Dratz EA, Granger BL, et al. (1997) The ligand binding site of the formyl peptide receptor maps in the transmembrane region. J Immunol 159: 4045–4054.
101. Davidson JS, Assefa D, Pawson A, Davies P, Hapgood J, et al. (1997) Irreversible activation of the gonadotropin-releasing hormone receptor by photoaffinity cross-linking: localization of attachment site to Cys residue in N-terminal segment. Biochemistry 36: 12881–12889.
102. Xie W, Jiang H, Wu Y, Wu D (1997) Two basic amino acids in the second inner loop of the interleukin-8 receptor are essential for Galpha16 coupling. J Biol Chem 272: 24948–24951.
103. Samson M, LaRosa G, Libert F, Paindavoine P, Detheux M, et al. (1997) The second extracellular loop of CCR5 is the major determinant of ligand specificity. J Biol Chem 272: 24934–24941.
104. Brelot A, Heveker N, Pleskoff O, Sol N, Alizon M (1997) Role of the first and third extracellular domains of CXCR-4 in human immunodeficiency virus coreceptor activity. J Virol 71: 4744–4751.
105. Picard L, Wilkinson DA, McKnight A, Gray PW, Hoxie JA, et al. (1997) Role of the amino-terminal extracellular domain of CXCR-4 in human immunodeficiency virus type 1 entry. Virology 231: 105–111.
106. Rucker J, Samson M, Doranz BJ, Libert F, Berson JF, et al. (1996) Regions in beta-chemokine receptors CCR5 and CCR2b that determine HIV-1 cofactor specificity. Cell 87: 437–446.
107. Damaj BB, McColl SR, Neote K, Songqing N, Ogborn KT, et al. (1996) Identification of G-protein binding sites of the human interleukin-8 receptors by functional mapping of the intracellular loops. FASEB J 10: 1426–1434.
108. Chiang N, Kan WM, Tai HH (1996) Site-directed mutagenesis of cysteinyl and serine residues of human thromboxane A2 receptor in insect cells. Arch Biochem Biophys 334: 9–17.
109. D'Angelo DD, Eubank JJ, Davis MG, Dorn GW (1996) Mutagenic analysis of platelet thromboxane receptor cysteines. Roles in ligand binding and receptor-effector coupling. *J Biol Chem* 271: 6233–6240.
110. Leong SR, Kabakoff RC, Hebert CA (1994) Complete mutagenesis of the extracellular domain of interleukin-8 (IL-8) type A receptor identifies charged residues mediating IL-8 binding and signal transduction. J Biol Chem 269: 19343–19348.
111. Hebert CA, Chuntharapai A, Smith M, Colby T, Kim J, et al. (1993) Partial functional mapping of the human interleukin-8 type A receptor. Identification of a major ligand binding domain. J Biol Chem 268: 18549–18553.
112. Ladoux A, Frelin C (2000) Coordinated up-regulation by hypoxia of adrenomedullin and one of its putative receptors (RDC-1) in cells of the rat blood-brain barrier. J Biol Chem 275: 39914–39919.
113. Fredriksson R, Lagerstrom MC, Lundin LG, Schioth HB (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol Pharmacol 63: 1256–1272.
114. Vaidehi N, Floriano WB, Trabanino R, Hall SE, Freddolino P, et al. (2002) Prediction of structure and function of G protein-coupled receptors. Proc Natl Acad Sci U S A 99: 12622–12627.
115. Bindewald E, Skolnick J (2005) A scoring function for docking ligands to low-resolution protein structures. J Comp Chem 26: 374–383.
116. Zhang Y, Kihara D, Skolnick J (2002) Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 48: 192–201.
117. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195–202.
118. Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL III (2000) Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 41: 86–97.
119. Reva BA, Finkelstein AV, Skolnick J (1998) What is the probability of a chance prediction of a protein structure with an RMSD of 6 A? Fold Des 3: 141–147.
120. Sayle RA, Milner-White EJ (1995) RASMOL: Biomolecular graphics for all. Trends Biochem Sci 20: 374.