Published in final edited form as:
Methods Mol Biol. 2012 ; 857: 231–257. doi:10.1007/978-1-61779-588-6_10.

Methods of protein structure comparison

Irina Kufareva[1] and **Ruben Abagyan**[1,2,*]

[1]UCSD Skaggs School of Pharmacy and Pharmaceutical Sciences, La Jolla, CA, 92093, USA

[2]San Diego Supercomputer Center, La Jolla, CA, 92093, USA

Abstract

Despite its apparent simplicity, the problem of quantifying the differences between two structures of the same protein or complex is non-trivial and continues evolving. In this chapter, we described several methods routinely used to compare computational models to experimental answers in several modeling assessments. The two major classes of measures, positional distance-based and contact-based, were presented, compared and analyzed.

The most popular measure of the first class, the global RMSD, is shown to be the least representative of the degree of structural similarity because it is dominated by the largest error. Several distance-dependent algorithms designed to attenuate the drawbacks of RMSD are described. Measures of the second class, contact-based, are shown to be more robust and relevant. We also illustrate the importance of using combined measures, utility-based measures, and the role of the distributions derived from the pairs of experimental structures in interpreting the results.

Keywords

protein structure comparison; modeling; docking; accuracy; assessment; root mean square deviation; atomic contacts; residue contacts; naïve model; Z-score; cumulative distribution function; VLS enrichment

1 Introduction

Applications of protein structures comparison methods

The majority of the proteome is made by amino-acid sequences that, due to evolutionary selection, reliably and reproducibly form essentially the same three-dimensional structure. This observation formed a basis of the “one sequence – one structure” paradigm that dominated the protein science for a long time. However, the growing redundancy of protein structure databases, i.e. the increase in the number of structures per protein (1-3), made it clear that these fascinating molecules possess a lot more than a simple, unique rigid structure, and that varying degrees of the inherent flexibility of proteins are critical for their functioning. Consequently, quantifying the structural differences in a sensible way becomes essential.

*Corresponding author: Ruben Abagyan, PhD, University of California, San Diego, Skaggs School of Pharmacy and Pharmaceutical Sciences, 9500 Gilman Drive, La Jolla, CA 92093, Phone: 1-858-822-3404, fax: 1-858-822-5591, rabagyan@ucsd.edu.

Structure comparison methods have been actively developed and used in the field of computational modeling assessments for quantitative evaluation of correctness of predicted models. Since 1994, a community-wide experiment called CASP (Critical Assessment of techniques for protein Structure Prediction, (4)) provides the modeling community with the possibility to evaluate their methods in blind prediction of structures of newly solved (but unpublished at the moment of the assessment) proteins. The submitted models are compared to an experimental structure using various criteria specifically developed for this task (5). In the recent years, other initiatives of this kind have emerged, including CAPRI (Critical Assessment of PRedicted Interactions (6)) and GPCR Dock (7), the assessment of modeling and docking methods for human G-protein coupled receptor targets, and the assessment of the docking and scoring algorithms (8, 9).

Despite the fact that the methods presented in this chapter were originally developed for comparison of computational models to the experimental answers, their applicability is not limited by the modeling assessments. They now find their use in identification, evaluation, understanding, and prediction of protein conformational changes which constitute the fundamental basis of their biological functioning.

Properties of an ideal protein similarity measure

An ideal measure should allow both a single ‘summary’ number within a fixed range (e.g. 0% to 100%) and an underlying detailed vector or matrix representation. The single number must distinguish well between related (correct) and non-related (incorrect) structure pairs, i.e. its distributions on the two sets must overlap to a minimal possible degree. It has to be relevant, i.e. capture the nature of protein folding or protein interaction determinants rather than satisfy simple geometric criteria. It has to have the minimal number of parameters, which in turn need to be well justified and understandable. It has to be stable and robust against minor or fractional (affecting a small fraction of the model) experimental and modeling errors; such changes in the structures should not lead to major leaps in the calculated similarity measure values. It has to capture the similarities or differences between the structures at any given level of accuracy/resolution. Ideally, it should have an intuitive visual interpretation. Although the complex nature of the problem prevents a universally acceptable single solution, some consensus measures are definitely emerging.

Characterization of protein structure comparison measures on protein structure pair datasets

In this chapter, we present an overview of several superimposition (distance)-based and contact based measures and characterize them by calculating their distribution on three sets of protein structure pairs. The first set consists of 130000 pairs of experimentally determined structures of identical proteins in PDB. It includes molecules related by non-crystallographic symmetry, structures determined in different crystal forms and in composition with different protein or small-molecule binding partners. The second and third sets are made of models of two G-protein coupled receptors, dopamine D3 receptor and chemokine receptor CXCR4, generated by participants of the community-wide GPCR Dock assessment (10) in summer 2010 prior to release of the experimental coordinates of these receptors in complex with small molecule (D3 and CXCR4) and peptide (CXCR4) modulators (11, 12). The second

and third sets are representative of the average modeling accuracy that can be achieved when the experimentally determined structures of closely related homologs are available (~40% of sequence identity as in the case of dopamine receptor D3 with previously solved β_1 and β_2 adrenergic receptors) or when the homology with existing structures is more distant (~25% of sequence identity as in the case of chemokine receptor CXCR4).

2 Methods

2.1 Main types of comparison measures

Sequence-dependent vs. sequence-independent methods—Sequence-dependent methods of protein structure comparison assume strict one-to-one correspondence between target and model residues. In sequence-independent methods, structural superimposition is performed independently, followed by the evaluation of residue correspondence obtained from such superimposition. The usefulness of the sequence-independent approach is limited to cases where a model approximately captures the correct target fold but the amino-acid sequence threading within this fold is incorrect, e.g. when one turn shift of an alpha-helix occurs. An example of an alignment-independent measure is the AL0 score routinely used in CASP model evaluation (13). AL0 score measures model accuracy by counting the number of correctly aligned residues in the sequence-independent superposition of the model and the reference target structure. A model residue is considered to be correctly aligned if the Ca atom falls within 3.8 Å of the corresponding atom in the experimental structure, and there is no other experimental structure Ca atom nearer. AL0 score values are clearly dependent on the superimposition; in its original implementation used for CASP model assessment, the score is calculated using the so called Local/Global Alignment (LGA (14)) superimposition of the two structures. A variety of sequence-independent structural alignment methods have been developed in the field: CE (15), DALI (16), DejaVu (17), MAMMOTH (18), Structal (19), FOLDMINER (20), KENOBI/K2 (21), LSQMAN (22), Matras (23, 24), PrISM (25), ProSup (26), SSM (27), and others.

The results of alignment-dependent and alignment-independent structure comparison are highly correlated with the exception of very distant homology cases. For the rest of this chapter, we will therefore focus on alignment-dependent methods of protein structure comparison and methods of identification of subtle similarities and differences between the models and the reference structures in rather accurate modeling applications.

Evaluation of local vs. global similarity—Identification of global versus local similarity represents two orthogonal directions in comparison of protein structures, i.e. structures that are most similar globally may not be the best in terms of local similarity. Flexible or disordered fragments such as long loops and/or termini are often poorly predicted and may significantly compromise the otherwise good similarity between structures. Relative domain movements observed in multi-domain proteins can also contribute to the poor global similarity scores. Focusing on local similarity helps to avoid these issues. Local similarity can be interpreted as a cumulative similarity score for all regions of the protein or, otherwise, can focus on a specific region such as, for example, ligand binding pocket, while ignoring the remaining parts of the protein.

Superimposition-based vs. superimposition-independent methods—Any method that relies on distance measurements between reference points in the model and their respective counterparts in the reference template requires prior superimposition of the model onto template, with the results of the comparison clearly dependent on the superimposition. Finding an optimal superimposition is an ambiguous task that has multiple solutions optimizing specific parameters, therefore, all superimposition dependent methods suffer from this ambiguity. A superimposition that minimizes the global Root Mean Square Deviation (RMSD) of the model to the template may not necessarily be the best solution for the reasons described above: such superimposition is often compromised by a small number of significantly deviating fragments. Superimposition of a specific subset may not resolve this issue because the choice of the subset is subjective and ambiguous. A method that iteratively optimizes the superimposition of two protein structures by assigning lower weights to most deviating fragments and, in this way by finding the largest superimposable core of the two proteins, is described below. However, even in this approach, the choice of weight decay function is rather arbitrary and subjective which may lead to multiple solutions introducing ambiguity in any similarity score derived from these superimpositions. Superimposition-independent methods, such as contact based measures are devoid of this ambiguity.

2.2 Distance-based measures of protein structure similarity

Root Mean Square Deviation (RMSD) is the most commonly used quantitative measure of the similarity between two superimposed atomic coordinates. RMSD values are presented in Å and calculated by

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}$$

where the averaging is performed over the n pairs of equivalent atoms and d_i is the distance between the two atoms in the i-th pair. RMSD can be calculated for any type and subset of atoms; for example, Cα atoms of the entire protein, Cα atoms of all residues in a specific subset (e.g. the transmembrane helices, binding pocket, or a loop), all heavy atoms of a specific subset of residues, or all heavy atoms in a small-molecule ligands.

The main disadvantage of the RMSD lies in the fact that it is dominated by the amplitudes of errors. Two structures that are identical with the exception of a position of a single loop or a flexible terminus typically have a large global backbone RMSD and cannot be effectively superimposed by any algorithm that optimizes the global RMSD. An example of such a pair is given by the active and inactive conformations of an estrogen receptor α (ERα) which are only different by the movement of a single helix 12 (Figure 1). By global backbone RMSD, this pair is virtually indistinguishable from the pair of albumin structures where multiple smaller scale rearrangements occur. The colored map in Figure 1 shows the distribution of the protein backbone RMSD for a large number of experimentally determined structure pairs of identical proteins in the PDB. It demonstrates that for the majority of pairs, the RMSD

ranges from 0 to 1.2 Å, due to inherent protein flexibility and experimental resolution limits. Figure 1 also presents the results of comparison of most accurate GPCR Dock 2010 models to their respective reference (answer) structures. It is clear that the backbone RMSD values are distributed around 2.3 Å for the easier homology modeling case, D3, and around 4.5 Å for the distant homology modeling case, CXCR4. It is however important to realize that these RMSD distributions do not reflect the true model accuracy because they are largely affected by flexible and poorly defined regions such as C-termini and extracellular loops in both GPCRs.

An important extension of the RMSD measure, the weighted RMSD (wRMSD), allows focusing on selected atomic subsets, for example, downplaying the regions known to be inherently unstructured:

$$wRMSD = \sqrt{\frac{\sum_{i=1}^{n} w_i d_i^2}{\sum_{i=1}^{n} w_i}}$$

Internal symmetry, ambiguities and RMSD—Any kind of RMSD-based measurement requires prior assignment of atom correspondences. In the case of Cα RMSD between two structures of the same protein, atom pair correspondence is established trivially via sequence alignment, however, measuring all heavy atom RMSD usually requires careful consideration of internal symmetry: the atom pair correspondence in such cases cannot be established unambiguously because some atoms within each structure are topologically equivalent to one another. For example, $C_{\delta 1}$ and $C_{\delta 2}$ atoms in a single phenylalanine (Phe) residue are topologically equivalent and therefore can be mapped into $C_{\delta 1}$ and $C_{\delta 2}$ atoms of the corresponding Phe residue of a different structure in two ways. The list of residues that cannot be mapped unambiguously includes: Arg, Asp, Glu, Leu, Phe, Tyr, and Val. Fortunately, the complexity of finding the optimal correspondence minimizing the overall side-chain RMSD is linear with respect to the number of residues. Figure 2 illustrates the distribution of heavy atom RMSD for a large set of small-ligand binding pocket pairs in PDB, calculated with and without side chain rotamer enumeration. While on average, finding the optimal atom correspondence reduces pocket RMSD by less than 0.1 Å, this effect is largely unpredictable and for extreme cases, can reach 0.5 Å.

RMS of dihedral angles—An approach complementary to Cartesian backbone RMSD is based on the representation of the protein structure in the internal coordinates that include bond lengths, planar bond angles, and dihedral torsion angles. For example, the geometry of a polypeptide chain backbone is described by the set of pairs of dihedral angle values, φ and ψ, which provides a way for a superimposition-independent structure comparison with the dihedral angle RMS used as the similarity scoring function. The dihedral angle RMS is complementary to the atom RMSD in the sense that it captures a different, less intuitive aspect of protein structure similarity. Modification of a small number of backbone dihedral angles can distort the global structure and packing beyond recognition while having only

marginal effect on the dihedral angle RMS. At the same time, very similar structures are sometimes characterized by significant variations in their dihedral angles simply because these variations may partially cancel each other (e.g. peptide flips). These phenomena are well illustrated by the experimental distribution of backbone dihedral angle RMSD as compared to backbone RMSD (Figure 1). For example, none of the three representative outlier clusters in terms of backbone RMSD, estrogen receptor, albumin, and myosin, is characterized by dihedral angle RMS deviating by more than two standard deviations from the experimental structure pair average. Similarly, the distribution of dihedral angle RMS of the GPCR Dock models to their respective reference structures lies in a region well populated by the experimental structure pairs, while their common sense similarity in terms of backbone RMSD remains on the margins of the experimental distribution.

Global Distance Test (GDT)—As described above, RMSD heavily depends on the precise superimposition of the two structures and is strongly affected by the most deviated fragments. A clever way to overcome both shortcomings was implemented in the two methods routinely used for CASP model evaluation, Global Distance Test (GDT) and Longest Continuous Segment (LCS) (28): here multiple superimpositions, each including the largest superimposable subset for one of the residues, are calculated between the two structures. In application to comparison of a model to an experimental answer, it means that for each residue from the model, the largest continuous (LCS) or arbitrary (GDT) set of the model residues is found that contains the residue and superimposes with the corresponding set in the reference structure under a selected RMSD (LCS) or distance (GDT) cutoff. The maximal residue set for each cutoff is chosen, followed by averaging over several fixed cutoffs (e.g. 1 Å, 2 Å, 4 Å, and 8 Å). The output of a GDT calculation represents a curve that plots the distance cutoff against the percent of residues that can be fitted under this distance cutoff. A larger area under the curve corresponds to more accurate prediction.

The distribution plot of GDT total score on the large set of experimental structure pairs is shown in Figure 3(a). Unlike the global backbone RMSD, the GDT measure recognizes structural similarity very well for the absolute majority of experimental pairs (GDT-TS > 50%). It is also more robust against small fragments movements. In particular, it effectively distinguishes the pair of active and inactive conformations of ERα which differ only in helix 12 conformation (GDT-TS = 93%), from the pair of albumin structures with multiple smaller scale domain distortions (GDT-TS = 60%).

TM-score—Another problem that one runs into when using RMSD to compare protein structures is that the RMSD distribution also depends on the size of the protein. This becomes important when the models of several different size proteins are evaluated in comparison with one another. The dependence of RMSD on the protein size can be eliminated by calculating the so-called TM-score (29):

$$TMscore = max\left[\frac{1}{L_{target}}\sum_{i=1}^{L_{aligned}}\frac{1}{1+\left(\frac{D_i}{D_0(L_{target})}\right)^2}\right]$$

Here L_{target} and $L_{aligned}$ are the number of residues in the reference structure and the aligned region of the model, respectively, and $D_0\left(L_{target}\right)=1.24^3\sqrt{L_{target}-15}-1.8$ is a distance scale derived from the analysis of large subsets of related and unrelated structures that is used to normalize the distances. Through its dependence on the L_{target}, the dependence of the obtained score on the target size is eliminated.

Iterative weighted superimposition and the associated superimposition error—The main CASP measure, GDT, is dependent on several arbitrarily chosen fixed distance cutoffs. This dependence is replaced by a continuous distance dependent weight in the iterative weighted superimposition algorithm (30). By unbiased weight assignment to different atomic subsets, this algorithm gradually finds the better superimposable core between the two structures. It includes the following steps:

1. The atomic equivalences are established between the two structures and a vector of per-atom weights $\{W_1, W_2, \ldots, W_n\}$ is set to $\{1, 1, \ldots, 1\}$.
2. The weighted superimposition is performed (31) and weighted RMSD is calculated as described above.
3. The deviations $\{d_1, d_2, \ldots, d_n\}$ are calculated for all atom pairs, and their X-quantile, d_X is determined. The quantile X is an input parameter for the procedure that defines the minimal size of the superimposable core to be found; by default it is equal to 50%.
4. The new weights are calculated according to the formula

$$W_i=exp\left(-d_i^2/d_x^2\right)$$

   The well superimposed atoms are assigned weights close to 1, while the weights associated with strongly deviating atom pairs get progressively smaller.
5. Steps from (2) through (4) are iterated until the weighted RMSD value stops improving *or* the specified maximum number of iterations is reached.

Following this superimposition, the similarity of the two structures can be evaluated by the weighted RMSD or by taking the average of weights recalculated for the structure according to step 4 with d_X set to a fixed value, e.g. 2 Å. The complement of this number, denoted *superimposition error* (E_{super}), ranges from 0 to 100% with lower values corresponding to more similar structure pairs:

$$E_{super}=100\%\times\left(1-\frac{1}{n}\sum_{i=1}^{n}exp\left(-\frac{d_i^2}{d_x^2}\right)\right)$$

The presence of a minority of strongly deviating atoms does not compromise the superimposition error while large discrepancies are accurately captured and quantified (Figure 4).

The algorithm resembles the one published by Damm and Carlson (32) with a few modifications, including the adaptable standard deviation for the Gaussian distribution (step 4 of the algorithm) and the way the weighted RMSD is calculated (normalization by the sum of weights. The adaptable denominator in the distribution ensures a better quality superposition.

Figure 3(a) represents the distribution of GDT-TS versus superimposition error for the experimental structure pair set and for the two sets of GPCR Dock 2010 models. The two measures are highly correlated. The adaptive nature of the GDT-TS measure that combined *multiple* superimpositions for different parts of poorly superimposable structures makes it more permissive; for the absolute majority of experimental structure pairs its value exceeds 50%. In contrast, superimposition error quantifies the structural deviations for a single weighted superimposition based on the largest common substructure; when two structures lack a significant common superimposable domain, superimposition error values may exceed 80%.

2.3 Contact-based measures of protein structure similarity

Contact-based measures rely on comparison of pairwise distances and/or interactions within one of the protein structures with the corresponding distances/interactions in the other structure rather than on finding the distances between the corresponding points in the two structures. They therefore possess the advantage of being superimposition-independent. Pairwise contact matrices found many applications as a method of 2D representation of 3D protein structure (33-36). Multiple possible contact definitions create a variety of contact-based protein structure similarity measures and make them adjustable to each particular subject area. In particular, by changing the contact distance cutoff one can make the contact-based similarity measure local or global to the evaluator's taste.

In general, a *contact* can be defined as an arbitrary continuous function of two points in a protein structure, not necessarily representing the true physical interaction of these points. Point selection defines the "grain" or resolution in the contact definition:

- Residue level contact measures

  - Coarse grain residue representation: single point per residue, e.g. Cα, Cβ, or a representative side chain "center of mass" point. Inter-residue contacts in the form of Cα-Cα distances were used, for example, in DALI algorithm for alignment-independent protein structure comparison (16).

  - Full atom residue representation

- Residue fragment level contact measures (same as above but each residue is divided into fragments, usually the backbone made of N, Cα, C, O atoms and the side chain)

- Atom level contact measures (contacts are calculated between the individual atom pairs)

Ways of determining the contact function include:

- Algebraic functions of inter-point distance (discontinuous, e.g. Heaviside step function, or continuous/smooth)
- Functions based on physical principles (contact surface area, interaction energy, etc.)
- Tabulated physics-based contact strengths as a function of inter-point distance and geometry.

Contact Area and Contact Strength Difference—In their Contact Area Difference (CAD) paper (37), Abagyan and Totrov came up with a contact definition that directly correlates with the strength of physical interactions, namely, they defined a residue contact as the difference in accessible surface area when calculated for a pair of residues separately or together. While this *contact area* measure provides the most realistic assessment of fold similarity between the two structures, it does require specific residue pairs to be in contact with about the same area. If the side chains are not packed correctly even with roughly similar fold, the distance will be large.

Contact functions based solely on C_α-C_α or C_β-C_β pairwise distances do not require correct (matching) residue-residue packing provided the backbones are similar. Given two residues whose C_α or C_β atoms are located at the distance of d Å, the residue *contact strength* can be calculated as

$$f(d) = \begin{cases} 1 & if \quad d<d_{min} \\ \frac{d_{max}-d}{d_{max}-d_{min}} & if \quad d_{min}<d<d_{max} \\ 0 & if \quad d>d_{max} \end{cases}$$

where d_{min} and d_{max} are predefined distance margin boundaries. The values of d_{min} and d_{max} can be chosen in such a way that the corresponding contact strengths are correlated with the pairwise residue contact areas which in turn describe the real physical residue interactions. C_β-C_β contacts approximate contact areas more accurately than C_α-C_α, because on average, C_β atoms are closer to the centers of mass of the residues they belong to. In (38), this approach was further improved by replacing C_β atoms by virtual points, C'_β, located in the direction of C_α-C_β bonds at the distance of $1.5 \times d(C_\alpha, C_\beta)$ from the C_α atom of each residue. This was shown to further improve the correlation between the calculated contact strengths and residue contact areas with the optimal margin boundaries found to be d_{min} = 4 Å and d_{max} = 8 Å.

When comparing two structures by their contacts, one builds two matrices of atomic contact strengths: $\boldsymbol{C}^R{}_{n\times n}$ for the first structure and $\boldsymbol{C}^M{}_{n\times n}$ for the second structure or model. The contact similarity matrix $\boldsymbol{C}^{R\cap M}$ is constructed using $\boldsymbol{C}^{R\cap M}[i,j] = \mathrm{Min}(\boldsymbol{C}^R[i,j], \boldsymbol{C}^M[i,j])$; its weight is found as $|\boldsymbol{C}^{R\cap M}| = \Sigma_{i,j}\boldsymbol{C}_{R\cap M}[i,j]$. This weight can be compared to one of three quantities: the weight of the reference contact matrix, $|\boldsymbol{C}^R|$, model contact matrix, $|\boldsymbol{C}^M|$, or the union of the two, $|\boldsymbol{C}^{R\cup M}|$, defined by $\boldsymbol{C}^{R\cup M}[i,j] = \mathrm{Max}(\boldsymbol{C}^R[i,j], \boldsymbol{C}^M[i,j])$ or $\boldsymbol{C}^{R\cup M}[i,j] = (\boldsymbol{C}^R[i,j] + \boldsymbol{C}^M[i,j])/2$. The three approaches result in quantities ranging from 0 to 100% and reflecting recall, precision, and accuracy with which the model reproduces the reference

structure contacts. Alternatively, one may choose to report the contact *differences* which simply complement the above similarity measures to 1 or 100% (contact distance or difference = 1. – contact similarity).

Figure 3(b) shows that for a large subset of PDB structure pairs, as well as for GPCR Dock 2010 models, contact strength differences calculated using the virtual C'_β points are highly correlated with contact area differences (CAD). For most pairs of experimentally determined structures of the same protein, protein flexibility and experimental errors lead to the contact strength differences of 5-20%. Small flexible fragments or even large domain movements have only minor effect on the contact strength matrices making the contact strength measures robust to elastic large scale deformations. At the same time, these measures are sensitive to major changes in packing occurring as a result of modeling errors: the best GPCR Dock models appear to be about 30% different from the reference structure in the case of D3 and about 40% different in the case of CXCR4.

Further developments of contact strength definitions may include their parameterization according to the interacting residue types, complementation of the C_β-C_β distances with other parameters to better capture the dependence of the contact strength or likelihood on the relative residue orientation, and elimination of the trivial contacts occurring due to the covalent linkages between the neighboring residues. These research topics are however beyond the scope of the present chapter.

The importance of multiple criteria analysis—The location of computational model populations on the plots of distance-based and contact-based measures of protein similarity in Figures 3(a) and 3(b) shows that in both cases, the models occupy the outskirts of the experimental distribution, with models built by closer homology (D3) being more accurate than distant homology models (CXCR4). The biggest insight however is gained when distance-based and contact-based measures are plotted against one another (Figure 3(c)). In these coordinates, it becomes clear that for the experimental structure, pairs may often differ in conformation (as reflected by superimposition error) or in contacts (as reflected by contact strength difference), but rarely in both. In contrast, computational models differ from their respective answers by both parameters simultaneously, especially in the more difficult modeling case of CXCR4. This observation stressed the importance of applying complementary structure similarity measures that combine distance-based and contact-based approaches.

2.4 Comparing protein-protein and protein-ligand complexes

Protein structure similarity measures presented above had the goal of comparing two structures of a single protein; however, the same general principles apply to evaluation of the predictions of molecular *interactions*. In 2002, the CAPRI (Critical Assessment of Predicted Interactions) experiment started with the focus on protein docking (39). Other initiatives followed including the GPCR Dock assessment started in 2008 and focused on small molecule docking to GPCR targets (7) as well as the recent assessment of ligand docking and virtual screening organized by Open-Eye (8, 9).

The task of *docking* is defined as prediction of the geometry and interactions in a *complex* of the given protein with either another protein (*protein docking*) or a small-molecule ligand (*small molecule docking*). In its pure form, the docking problem is based on the assumption that the structures of the unbound components are available. However, in real-life applications it is rarely the case; even when such structures do exist, they may not be directly usable for complex geometry prediction because of the induced fit effect (40, 41) and uncertainties in amino acid tautomerization, protonation, and hydration (42). If the unbound structures do not exist they must be generated by homology for proteins and by 2D to 3D conversion for small molecules which introduces an additional level of difficulty in the docking protocol.

Methods that are used for the evaluation of docking predictions are largely based on the same principles as the methods of comparison of protein structures described above. However because the focus is on the intermolecular interactions, one must ensure that the unrelated discrepancies in the structures of the individual interaction partners have minimal effect on the evaluation outcome.

Let us assume for simplicity that the complex of interest consists of only two molecules, and that one of them (a protein) can be classified as a receptor while the other one (another protein, a peptide, or a small molecule) is a ligand. In protein-protein complex prediction, the designation of one of the partners as a receptor is rather arbitrary and may be performed based on the size, rigidity, availability of structural information or other criteria.

The most common way to evaluate the correctness of the docking geometry is to measure the Root Mean Square Deviation (RMSD) of the ligand from its reference position in the answer complex after the optimal superimposition of the receptor molecules. The choice of this optimal superimposition is the first subjective decision that the evaluator must make, especially in the case when the receptor had to be modeled and therefore the reference and the modeled receptor structures are significantly different. To reduce the effect of the irrelevant incorrectly modeled receptor parts, it is important that the receptor superimposition is performed by a smaller subset of atoms that includes the immediate binding interface (or binding pocket in case of a small molecule docking problem). Criteria for the selection of the binding interface residues should be carefully formulated and stated upfront; the usual procedure involves selection of residues located at a certain distance from the ligand in the reference structure followed by expansion of this selection through the sequence so that the short discontinuous stretches of residues are either merged or eliminated. The final selection must consist of continuous sequence stretches of at least 4-5 residues each to ensure that they can be properly aligned between the model and the reference structure. The interface selection must be derived from the reference structure and propagated to each complex model by the alignment-derived residue correspondence.

The interface atoms or pocket residues must now be superimposed for each model onto the reference structure. While the standard superimposition approach is the optimization of the selection heavy atom RMSD, flexible side-chains, loops, and termini may compromise the superimposition quality and therefore one of the more robust superimposition methods described above is preferred. Once the superimposition is performed, the time comes to

measure the RMSD between the ligand atoms in the model and the reference structures. The spectrum of caveats and challenges here is similar to that described in the previous paragraphs about RMSD, with the important distinction that whether the atoms in direct contact with the receptor constitute a minor or a major part of the ligand, they should remain the primary focus of the RMSD calculation. On the contrary, parts of the ligand distant from the interface or not in direct contact with the receptor must be down-weighted or disregarded in such an evaluation. For example the contribution of the solvent-exposed parts of the ligand to the overall similarity score was eliminated in the GPCR Dock 2008 assessment (the solvent exposed phenoxy group of the adenosine A_{2A} receptor antagonist (7, 43, 44) (Figure 5 (a)). In protein docking, elimination of the effects of ligand parts not directly involved in the interaction with the receptor becomes critical (Figure 5(b)).

Due to these caveats and ambiguities, positional distance-based measures need to be complemented with the contact measures of docking complexes. Contact definitions for protein-protein complexes are identical to the single protein case but are applied to intermolecular residue contacts only. Contacts are calculated between each pair of residues in the receptor and in the ligand and can involve C_α-C_α, C_β-C_β, virtual C'_β-C'_β distances as well as the actual residue contact areas. In case of small molecule ligands, because the scope of the problem is smaller and because atomic-level interactions become primarily important, the definition of contact strengths should be extended to allow calculation of the inter-atomic instead of the inter-residue contacts.

The definition of an atomic contact used for scoring protein-ligand complexes in the GPCR Dock 2008 modeling and docking assessment (7) involved a step-wise function of interatomic distance equal to 1 below the specified distance cutoff (4 Å) and 0 otherwise (Figure 6(a), black curve). In other words, each of the models was characterized by the set of all ligand-receptor atom pairs located at the distance of ≤ 4 Å; this set was compared to the corresponding atom pair set in the reference structure (45). While simple conceptually and computationally, this “hard distance cutoff” approach leads to unstable and discontinuous behavior of the contact difference function, because minor changes in the ligand and side-chain conformation may result in large leaps in the number of matching contacts (Figure 6(b)). To avoid this problem, the ligand-receptor atomic contact definition was refined in GPCR Dock 2010 with the *continuous decrease margin* approach in the spirit of (38). Instead of abruptly dropping to zero at the single cutoff of d_0, the contact strength gradually decreased between two distances, d_{min} and $d_{max} = d_{min} + m$, where m is the margin size. The margin boundaries, d_{min} and d_{max}, were adjusted so that the average number of contacts calculated with and without the margin is the same using the following equation:

$$d_{min}=d_0 - r \times m; \quad d_{max}=d_0+r\,(1-r) \times m$$

where r was calculated as $r = 0.49 + 0.17 \times m/d_0$. This equation was obtained by linear regression on the large number of complex structures.

The atomic contact definition can be further improved by making it atom-type dependent and/or orientation dependent; this will allow, for example, automatic assignment of higher weight to correctly predicted hydrogen bonds between the ligand and the protein.

Interatomic contact strength matrices can be calculated for the model and the reference structure. Taking the element-wise minima produces the matrix of correctly identified contact strengths which can be further compared to the reference matrix to give contact recall, model matrix for contact precision, or a combination of the two to give some form of contact accuracy. In cases where the physical atom-atom contacts are measured, contact precision can usually be disregarded: molecular geometry and van der Waals interactions impose natural constraints onto precision values because they limit the number of physical contacts that can be made.

The phenomenon of internal molecular symmetry may become a serious hurdle for evaluation of similarity of a predicted docking complex to the experimentally derived answer by either distance-based or contact-based measures. If the ligand possesses any symmetrical groups, all topologically equivalent mappings of its atom set onto itself must be considered. For example, because the resonance-stabilized thiol form of the thiourea group is symmetric, as many as 16 atom permutations in the compound IT1t (Figure 6(c)) result in exactly the same ligand covalent geometry and bond topology; all of these have to be tested when determining either RMSD or contact similarity of this compound to its copy in a different structure. In combination with the internal symmetry of neighboring side-chains, this may easily lead to exponential growth of computational complexity.

2.5 Combining measures for ranking a model population

As described above, the concept of protein structure similarity involves multiple criteria leading to a very different ranking of models. Combining these criteria into a single numerical score seeks a fair balance between complimentary measures each representing an important part of the whole picture. However, the uncertainties of this combination (which terms to use and now to normalize them) often create even more confusion. An approach that is routinely used in CASP is based on the analysis of the distribution of scores calculated for each individual assessment criterion and each individual modeling target. Score mean and standard deviation (SD) are calculated for each criterion after which the score is converted into the *intra-population* Z-score by taking

$$Z_S = (S - \mu_S) / \sigma_S$$

where μ_S and σ_S are the average and standard deviation of the score S. Z-scores can be easily modified so that a larger value corresponds to a higher level of accuracy. In many cases it is beneficial to remove the lowest accuracy outliers in the set so that they do not significantly affect the overall distribution. The intra-population Z-scores calculated in this way for the multiple assessment criteria (e.g. RMSD and contacts) are then averaged to obtain a single Z-score that is used to rank the models for the given target.

The intra-population Z-score approach allows bringing multiple differentially distributed criteria onto the same scale. In this way, it enables a fair comparison of the models for a given target protein without giving preferences to any of the assessment criteria and provides a way to determine the most accurate models in the population. The approach however is not devoid of drawbacks. Most importantly, it gives no information about *how* accurate the most accurate models are; therefore, Z-scores appear incomparable between different targets of varying difficulty. For a challenging target, even a model with the highest Z-scores is often extremely far from truth, while for targets with closer homology to the existing templates lower Z-score values may correspond to very accurate predictions. Furthermore, the choice of measures to be included in the Z-score is not only subjective, but often is decided only at the evaluation stage. Combining correlated criteria implicitly gives them higher weight in the overall Z-score. Finally, because not all assessment criteria are normally distributed, conversion of these values into Z-scores creates somewhat distorted statistics, in this case probabilities (a.k.a., the p-values) or their logarithms calculated for specific distributions make better contributions to the score (however they cannot be mixed with the Z-scores).

The main problem of the intra-population Z-score approach is the absence of information about how close the models are to the correct answer. Even within a population of completely incorrect models some model will be the “best”. To overcome this problem a better method is to compare the predictions with the distribution of the natural structural differences between “correct”, i.e. experimentally determined structure pairs. With the wealth of protein structure information growing exponentially (1), it is easy to calculate, for example, the distribution of ligand RMSD values between multiple structures of the same complex. After that, one can normalize a model ligand RMSD value from the reference structure by determining what fraction of experimental structure pairs are characterized by the same or higher ligand RMSD (cumulative distribution function, CDF). In principle, it is possible to calculate the Z-score of each model in the reference experimental value distribution, however, caution is necessary for criteria with non-normal distributions. The flipside of the CDF approach is that in difficult cases the majority of the models may appear far too distant from the real target structure to receive a non-zero CDF score; therefore, the model population ranking may become impossible.

To illustrate the concept of CDF percentiles, we calculated their values for the sets of D3 and CXCR4 models in GPCR Dock 2010 (Table 1). For example, in comparison with the most favorable reference (answer) structure, an average model in the top half of the D3 set was better than 5.24% of experimental pairs by superimposition error, while an average CXCR4 model was only better than 1.68%. Unlike intra-population Z-scores, these CDF percentiles project the model quality on the uniform scale of correctness which makes them comparable not only (i) between the models, but also between (ii) different targets and (iii) assessment criteria. For example, by averaging CDF percentiles over the four comparison criteria in Table 1, we can obtain the CDF score of 3.33% for an average D3 model but only 0.86% for an average CXCR4 model, which is representative of both absolute and relative accuracy of the modeling in the two cases. This result is of course expected given the fact that closer homology modeling templates were available in PDB for D3 than for CXCR4 at

the time of the assessment. It is quite encouraging however, that several D3 predictions fell into a significantly populated region of the experimental distribution, with the most accurate D3 model achieving 17.57% and 13.46% CDF values in terms of ligand RMSD and contacts, respectively.

3 Notes

3.1 X-ray structures as “golden standard” in model evaluation

Structural variability within sets of protein structures determined for the same parent protein but in different crystal or molecular environments has been acknowledged and quantified in several publications (3, 30, 46). On one hand, such variability may be due to the inherent protein flexibility triggered by a different complex composition or crystallization environment. On the other hand, it may be an artifact of the limited resolution of the structure determination techniques and the inevitable experimental errors. The extent of conformational changes observed between multiple structures of the same protein ranges from minor side-chain rearrangements to large-scale domain and loop movements, and depends on the protein functional class, crystal form and contacts (47), co-crystallized interaction partners (30) and other factors. A large-scale analysis of a redundant set of protein structures was performed in (3) and led the authors to the conclusion about the limited possibility of modeling proteins with multiple conformational states. In this regard, a legitimate question is whether a set of crystallographic coordinates represents an undisputable truth about native, biologically relevant structure of the protein, and whether it is conceptually correct to judge models by the degree of their structural similarity to the X-ray “answer”. The question is open-ended, because up to date, X-ray crystallography is the only experimental method capable of elucidating proteins and their interactions at the atomic resolution level. Using crystallographic structures as modeling standards is therefore inevitable; however, several measures can be taken to account for arising issues:

- Compare the model to the relevant conformational states and complex compositions
- Compare the model to the conformational ensemble and not a single structure (choose either the best or the average score)
- Down-weight or eliminate the contribution of flexible or poorly defined regions
- Report comparison scores in context of their distribution between the multiple structures in the ensemble.

These steps help to translate the knowledge about the “natural” protein variation into an improved comparison measure. For example, in GPCR Dock 2010, all dopamine D3 receptor models were compared to the two non-crystallographic symmetry related complexes in the reference structure, PDB 3pbl. The CXCR4 models were compared to the ensemble of as many as 8 reference complexes. For each combination of criteria, the values were reported in comparison with the most favorable reference in this ensemble. Moreover, the primary focus of the assessment was made on prediction of the ligand binding area and interactions which, in contrast to the intracellular or extracellular loops, are unlikely to be significantly affected by protein flexibility.

3.2 Separating trivial from non-trivial: the naïve models

In addition to the question of how close a model is to the experimental structure, it is also important to know how far it is from the result of applying a sensible but trivial procedure. The so called “naïve” models allow evaluation of the contribution of newly developed advanced modeling and refinement procedures in comparison with the most simple and straightforward approaches. In a way, the role of naïve models is similar to the role of placebo in drug clinical evaluation. Quite interestingly, the number of drugs that fail in the clinical trials by the reason of being no more effective than placebos constantly increases (48), leading some to the conclusion that the placebo effect is strengthening. Similarly, the constant method development in protein structure prediction makes the “naïve” models increasingly sophisticated thus shifting the baseline in model evaluation.

The most straightforward way to build a naïve model is threading the target sequence through a homology template without any subsequent optimization, or, in some cases, with fast side-chain optimization aimed at removal of major steric clashes. Even along this simple path, several factors may dramatically affect the quality and the degree of naivety of the models. They include (i) choice of the homologous protein and (ii) of the specific structure of that protein to be used as the homology template, as well as (iii) choice of the target-template sequence alignment which, with the exception of the extremely high homology cases, usually appears ambiguous. Figures 3(d,e,f) present the scatter plot of such naïve model on the background of the top half of GPCR Dock models. The accuracy range of the naïve models is substantial; in this case the range is primarily determined only by the choice of the homology template because we used our best knowledge sequence alignment in each case. For homology modeling, we used the six GPCR structures available in PDB prior to the 2010 GPCR Dock assessment: those of bovine rhodopsin in dark (bRho) and light-activated ligand free (opsin) states (49-51), β_1 and β_2 adrenergic receptors (52-54), and adenosine A_{2A} receptor (44). Our naïve models are close to the center of the distribution of the assessment models which may indicate the similarity of the approaches used by the GPCR Dock participants. However, a few models stand out and fall closer to the natural variation zone.

Whenever the modeling process includes not only modeling of the protein structure but also the docking of a protein or a small-molecule ligand, the definition of a naive model becomes even less defined. In rare cases when a homologous complex structure exists, it may be used to build a naïve, non-optimized model of the target complex as long as the target and the template ligands can be unambiguously (structurally) aligned. For protein ligands, the alignment may be based on sequence homology; but small molecules or in some cases short peptides may require finding the maximal common substructure between the target and template ligands, or establishing the correspondence in some other non-trivial way.

As an example, let us consider the challenges of building a “naive” model of the dopamine D3 receptor complex with eticlopride. This molecule belongs to a large class of aminergic antagonists and shares some degree of pharmacophoric similarity with previously crystallized antagonists of β_2 adrenergic receptor, carazolol and timolol. We performed pharmacophore-based alignment of the three-dimensional eticlopride molecule onto the structures of these two adrenergic antagonists. Because the procedure produced several

answers, the top ten chemical alignments were taken for each template, each was combined with the six naïve models generated by sequence threading and locally minimized to eliminate severe side-chain/ligand steric clashes. This produced a population of "naive" D3 complex models presented in Figure 7 (b). The accuracy range of these models is huge. Some of them approach (though none of them exceeds!) the level of accuracy of the best D3 models in GPCR Dock 2010. Though the step of scoring and selection was not employed in this exercise, it illustrates that (i) the level of model naivety may be highly variable, especially in case of protein-ligand docking complexes, and (ii) "naïve" sampling is capable of producing very accurate models.

In summary: the naïve models are useful to separate the actual advances from the trivial sensible approach; however, their definition appears too ambiguous to make them reliable standards of structure comparison and evaluation.

3.3 Evaluation of model quality without direct comparison to the reference structure

3.3.1 Spatial feasibility—The first question that has to be answered about a model is, in fact, not the degree of its similarity to the reference structure, but its spatial feasibility. This kind of evaluation is widely used to assess local errors in crystallographic coordinates during the refinement process or submissions for a modeling competition. The evaluation may be based on geometrical, stereochemical or statistical criteria, e.g. WhatCheck (55, 56), PROCHECK (57), or MolProbity (58), while some others, e.g. ICM Protein Health (59) use realistic normalized force field residue energies, where the expected distributions for the energies for each residue are derived from high quality crystal structures. An alternative approach involves the cumulative residue pseudo-energies or scores calculated as function of local atom, residue, secondary structure, accessibility environment and trained to predict the deviations from the near native-models. Multiple methods (VERIFY3D, PROSA, BALA, ANOLEA, PROVE, TUNE, REFINER, PROQRES) were integrated into a meta-server called MetaMQAP and trained to predict the residue deviations. While the individual residue predictions may not be accurate, combining different methods and averaging the residue signal in a five residue window led to impressive quality prediction values (60).

3.3.2 Ligand screening selectivity—Despite the obvious progress in protein structure prediction methodology and tools, the gain in modeling accuracy, as evaluated by similarity to the experimentally solved answer, has become less prominent in recent years (4). It appears therefore that the progress in the protein structure prediction area is reaching a certain plateau, and that the question of primary importance at this stage is not how to make models more similar to the experimentally derived structures, but how to make the most use of these models at the given level of prediction accuracy. Because one of major applications of modeling is in rational structure-based drug discovery and optimization, it appears relevant to directly evaluate the drug discovery potential of the models.

In the area of prediction of protein/ligand complex structures, Virtual Ligand Screening (VLS) enrichment by a model represents a clever way of evaluation of the model compliance with the experimental data in the form of small molecule chemical activity against the modeled protein. In this experiment, a large set of chemicals containing known

potent binders to the protein of interest (1-10% of the set) and diverse decoys of similar molecular weights and atom counts (90-99% of the set) is docked to the model, and the molecules are ranked by their predicted binding affinity. The model that efficiently and selectively scores the active molecules better than decoys apparently has a good potential for *de novo* drug discovery efforts. Quite interestingly, it appears also that such models often are most accurate in terms of predicted contacts between the ligand and the pocket atoms. For example, in both GPCR Dock 2008 (7) and GPCR Dock 2010 (10) assessments, model selection by VLS enrichment proved to be a successful strategy leading to most accurate predictions.

An important question is how to quantify VLS enrichment by a model. One of the traditional approaches to the problem involves calculation of the area under the so-called Receiver Operating Characteristic curve (ROC curve) which plots the ratio of true positives (TP, y-axis) against the ratio of false positives (FP, x-axis) in the top portion of the hit list ordered by the predicted binding affinity for each value of the affinity cutoff. A variation of the ROC curve is built when the fraction of TP is plotted against the total number of compounds scoring below the given cutoff rather than the FP rate. Both approaches suffer from the inability to distinguish early enrichment from late enrichment, and therefore are often complemented by the specific enrichment factors (EF) at the given FP rate, e.g. EF1 denotes the fraction of correct, active compounds that score better than 1% of the top-scoring decoys.

The normalized square-root Area Under Curve (nsAUC) is the area under the curve that plots the fraction of TP on top of the hit list (y-axis) against the *square root* of the total number of compounds scoring below the given cutoff (x-axis). Previous studies indicated that this measure is more representative of the true model selectivity than either the regular ROC which under-stresses the initial compound recognition (Figure 8) of the log-AUC (61, 62) which over-stresses it. With the non-normalized square-root AUC approach, the ideal sAUC (perfect recognition, all actives are ranked better than all inactives) and the random sAUC (actives are retrieved at the same rate as total compounds in the set, no recognition) are given by

$$sAUC_{ideal} = \frac{1}{c}\int_0^{\sqrt{c}} x^2 dx + \left(1 - \sqrt{c}\right) = \frac{1 - 2\sqrt{c}}{3}$$

and

$$sAUC_{rnd} = \int_0^1 x^2 dx = \frac{1}{3},$$

respectively. Here *c* is the fraction of the active compounds in the set. For the purpose of comparing the AUC across different datasets, sAUC is normalized to get:

$$nsAUC = \frac{sAUC - sAUC_{rnd}}{sAUC_{ideal} - sAUC_{rnd}} \times 100\%$$

that ranges from 0% (random) to 100% (ideal).

Finally, the VLS enrichment is not the only possible way to incorporate ligand binding information in the modeling process. Alternative approaches may be based on known active ligand pharmacophores, for example, by detection of complementarity of such pharmacophores to the model pocket. Though not directly measuring the drug discovery potential of the model, this approach also proved fruitful for increasing the accuracy of the GPCR-ligand complex structure prediction in GPCR Dock 2010 (10).

4 Acknowledgments

Authors wish to thank the organizers and the participants of the GPCR Dock 2010 assessment for providing the model statistics, Max Totrov and Eugene Raush for implementing some of the core functions in ICM, Manuel Rueda for helpful discussions and Karie Wright for help with manuscript preparation. We would like to acknowledge financial support by NIH, grants # R01 GM071872, U01 GM094612, and U54 GM094618.

5 References

1. Gabanyi M, Adams P, Arnold K, Bordoli L, Carter L, Flippen-Andersen J, Gifford L, Haas J, Kouranov A, McLaughlin W, et al. Journal of Structural and Functional Genomics. 2011:1–10. [PubMed: 21533787]
2. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, PrliÄ‡ A, Quesada M, Quinn GB, Westbrook JD, et al. Nucleic Acids Research. 2011; 39:D392–D401. [PubMed: 21036868]
3. Burra PV, Zhang Y, Godzik A, Stec B. Proceedings of the National Academy of Sciences. 2009; 106:10505–10510.
4. Kryshtafovych A, Fidelis K, Moult J. Proteins: Structure, Function, and Bioinformatics. 2009; 77:217–228.
5. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. Proteins: Structure, Function, and Bioinformatics. 2009; 77:18–28.
6. Wodak SJ. Proteins: Structure, Function, and Bioinformatics. 2007; 69:697–698.
7. Michino M, Abola E, participants GD, Brooks CL, Dixon JS, Moult J, Stevens RC. Nat Rev Drug Discov. 2009; 8:455–463. [PubMed: 19461661]
8. Warren, G.; Nevins, N.; McGaughey, G. 241st ACS National Meeting; Anaheim, CA. 2011.
9. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, et al. Journal of Medicinal Chemistry. 2005; 49:5912–5931. [PubMed: 17004707]
10. Kufareva I, Rueda M, Katritch V, participants GD, Stevens RC, Abagyan R. Structure submitted. 2011
11. Wu B, Chien EYT, Mol CD, Fenalti G, Liu W, Katritch V, Abagyan R, Brooun A, Wells P, Bi FC, et al. Science. 2010; 330:1066–1071. [PubMed: 20929726]
12. Chien EYT, Liu W, Zhao Q, Katritch V, Won Han G, Hanson MA, Shi L, Newman AH, Javitch JA, Cherezov V, et al. Science. 2010; 330:1091–1095. [PubMed: 21097933]
13. Kryshtafovych A, Venclovas Č , Fidelis K, Moult J. Proteins: Structure, Function, and Bioinformatics. 2005; 61:225–236.
14. Zemla A. Nucleic Acids Research. 2003; 31:3370–3374. [PubMed: 12824330]
15. Shindyalov IN, Bourne PE. Protein Engineering. 1998; 11:739–747. [PubMed: 9796821]
16. Holm L, Sander C. Journal of Molecular Biology. 1993; 233:123–138. [PubMed: 8377180]

17. Kleywegt, GJ.; Jones, AT. Methods in Enzymology. Academic Press; 1997. p. 525-545.
18. Ortiz AR, Strauss CEM, Olmea O. Protein Science. 2002; 11:2606–2621. [PubMed: 12381844]
19. Levitt M, Gerstein M. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95:5913–5920. [PubMed: 9600892]
20. Shapiro J, Brutlag D. Nucleic Acids Research. 2004; 32:W536–W541. [PubMed: 15215444]
21. Szustakowski JD, Weng Z. Proteins: Structure, Function, and Bioinformatics. 2000; 38:428–440.
22. Kleywegt GJ. Acta Crystallogr D Biol Crystallogr. 1996; 52:842–857. [PubMed: 15299650]
23. Kawabata T, Nishikawa K. Proteins. 2000; 41:108–122. [PubMed: 10944398]
24. Kawabata T. Nucleic Acids Res. 2003; 31:3367–3369. [PubMed: 12824329]
25. Yang A-S, Honig B. Journal of Molecular Biology. 2000; 301:665–678. [PubMed: 10966776]
26. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS. Protein Engineering. 2000; 13:745–752. [PubMed: 11161105]
27. Krissinel E, Henrick K. Acta Crystallographica Section D. 2004; 60:2256–2268.
28. Zemla A, Venclovas, Moult J, Fidelis K. Proteins Suppl. 2001; 5:13–21.
29. Zhang Y, Skolnick J. Proteins: Structure, Function, and Bioinformatics. 2004; 57:702–710.
30. Abagyan R, Kufareva I. Methods Mol Biol. 2009; 575:249–279. [PubMed: 19727619]
31. McLachlan AD. J Mol Biol. 1979; 128:49–79. [PubMed: 430571]
32. Damm KL, Carlson HA. Biophysical journal. 2006; 90:4558–4573. [PubMed: 16565070]
33. Phillips DC. Biochem Soc Symp. 1970; 30:11–28. [PubMed: 4923824]
34. Nishikawa K, Ooi T. J.Theor.Biol. 1974; 43:351–274. [PubMed: 4818352]
35. Liebman MN. Biophys. J. 1980; 32:213–215. [PubMed: 19431357]
36. Sippl MJ. Journal of Molecular Biology. 1982; 156:359–388. [PubMed: 7086905]
37. Abagyan RA, Totrov MM. J Mol Biol. 1997; 268:678–685. [PubMed: 9171291]
38. Marsden B, Abagyan R. Bioinformatics. 2004; 20:2333–2344. [PubMed: 15087320]
39. Lensink MF, Wodak SJ. Proteins: Structure, Function, and Bioinformatics. 2010; 78:3085–3095.
40. Bottegoni G, Kufareva I, Totrov M, Abagyan R. J Med Chem. 2009; 52:397–406. [PubMed: 19090659]
41. Totrov M, Abagyan R. Curr Opin Struct Biol. 2008
42. Coupez B, Lewis RA. Curr Med Chem. 2006; 13:2995–3003. [PubMed: 17073642]
43. Katritch V, Rueda M, Lam PC-H, Yeager M, Abagyan R. Proteins. 2010; 78:197–211. [PubMed: 20063437]
44. Jaakola V-P, Griffith MT, Hanson MA, Cherezov V, Chien EYT, Lane JR, Ijzerman AP, Stevens RC. Science. 2008; 322:1211–1217. [PubMed: 18832607]
45. Rueda M, Katritch V, Raush E, Abagyan R. Bioinformatics. 2010; 26:2784–2785. [PubMed: 20871105]
46. Stroud RM, Fauman EB. Protein Science. 1995; 4:2392–2404. [PubMed: 8563637]
47. Eyal E, Gerzon S, Potapov V, Edelman M, Sobolev V. Journal of Molecular Biology. 2005; 351:431–442. [PubMed: 16005885]
48. Golomb BA, Erickson LC, Koperski S, Sack D, Enkin M, Howick J. Annals of Internal Medicine. 2010; 153:532–535. [PubMed: 20956710]
49. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Trong IL, Teller DC, Okada T, Stenkamp RE, et al. Science. 2000; 289:739–745. [PubMed: 10926528]
50. Scheerer P, Park JH, Hildebrand PW, Kim YJ, Krausz N, Choe H-W, Hofmann KP, Ernst OP. Nature. 2008; 455:497–502. [PubMed: 18818650]
51. Park JH, Scheerer P, Hofmann KP, Choe H-W, Ernst OP. Nature. 2008; 454:183–187. [PubMed: 18563085]
52. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AGW, Tate CG, Schertler GFX. Nature. 2008; 454:486–491. [PubMed: 18594507]
53. Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi H-J, Yao X-J, Weis WI, Stevens RC, et al. Science. 2007; 318:1266–1273. [PubMed: 17962519]

54. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, Choi H-J, Kuhn P, Weis WI, Kobilka BK, et al. Science. 2007; 318:1258–1265. [PubMed: 17962520]
55. Hooft RW, Vriend G, Sander C, Abola EE. Nature. 1996; 381:272–272. [PubMed: 8692262]
56. Vriend G. J Mol Graph. 1990; 8:52–56. [PubMed: 2268628]
57. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. Journal of Applied Crystallography. 1993; 26:283–291.
58. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. Acta Crystallographica Section D. 2010; 66:12–21.
59. Maiorov V, Abagyan R. Fold Des. 1998; 3:259–269. [PubMed: 9710569]
60. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM. BMC Bioinformatics. 2008; 9:403–403. [PubMed: 18823532]
61. Jain A, Nicholls A. Journal of Computer-Aided Molecular Design. 2008; 22:133–139. [PubMed: 18338228]
62. Clark R, Webster-Clark D. Journal of Computer-Aided Molecular Design. 2008; 22:141–146. [PubMed: 18256892]

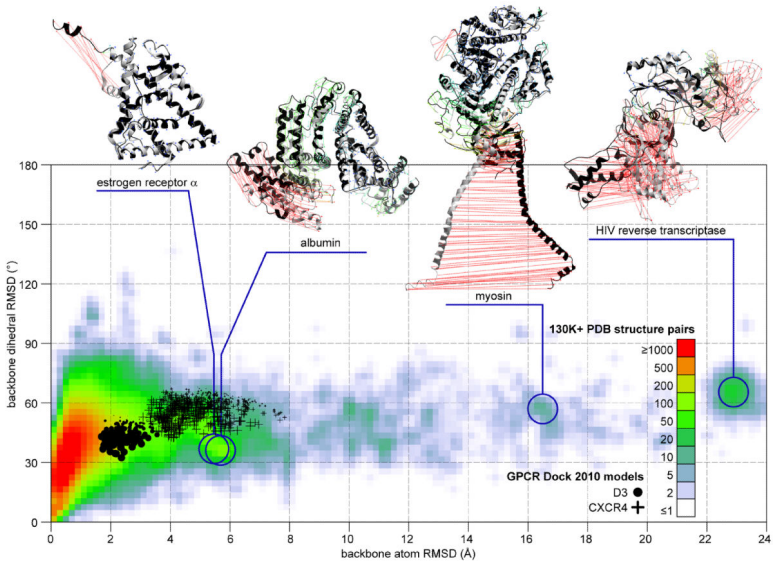

Figure 1.

Distribution of backbone atom RMSD/backbone dihedral RMSD values for a large number of experimentally determined pairs of protein structures in PDB. Representative structure pairs are shown. Computational models of dopamine D3 receptor (•) and chemokine receptor CXCR4 (+) are presented on the experimental structure pair background.

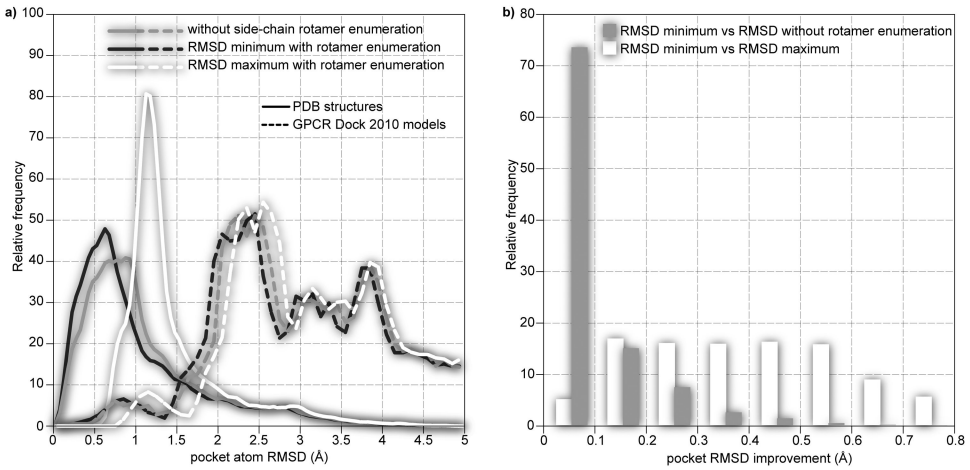

Figure 2.

Because some protein sidechains possess internal symmetry, atom correspondence between two identical sets of residues can be established in multiple ways resulting in different atom RMSD values. Finding the true pocket RMSD requires enumeration of possible correspondences between multiple equivalent side-chain rotamers. Equivalent rotamer enumeration lowers the calculated pocket RMSD by ~ 0.07Å on average, and by as much as 0.5 Å in extreme cases. Statistics collected from a set of 65000 PDB pocket pairs is presented as well as the results of analysis of GPCR Dock 2010 models.

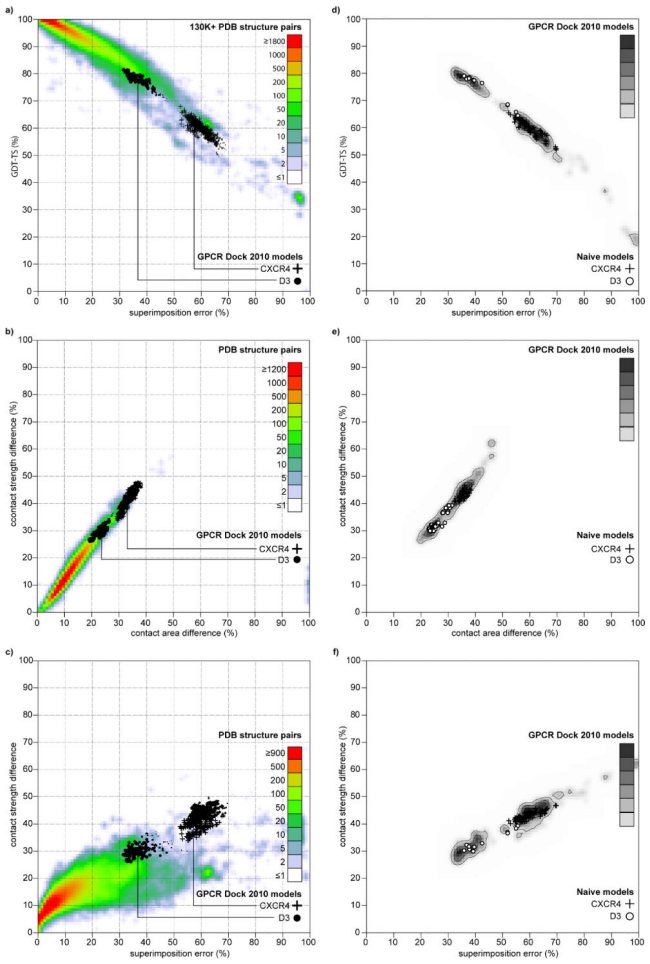

Figure 3.

Distribution of different measures of protein structure similarity for a set of 130000 protein structure pairs in PDB (**a-c**, heat map), GPCR Dock 2010 models (**a-c**, • for D3 and + for CXCR4; **d-f**, heat map), and naïve models of GPCR Dock 2010 targets (**d-f**, ○ for D3 and + for CXCR4). Only the top half of each GPCR Dock model set is shown: models less accurate than average are eliminated.

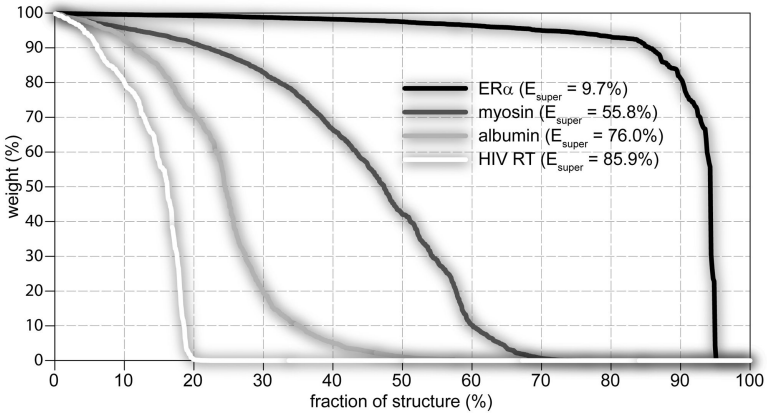

Figure 4.

Calculation of superimposition quality and superimposition error for representative structure pairs from Figure 1. Superimposition quality is calculated as the area under the weight curve; superimposition error (E_{super}) is its complement to 100%. Essentially identical structure pairs like active/inactive conformation pair of ERα receive high weight for the majority of the structure and, consequently, low value of superimposition error.

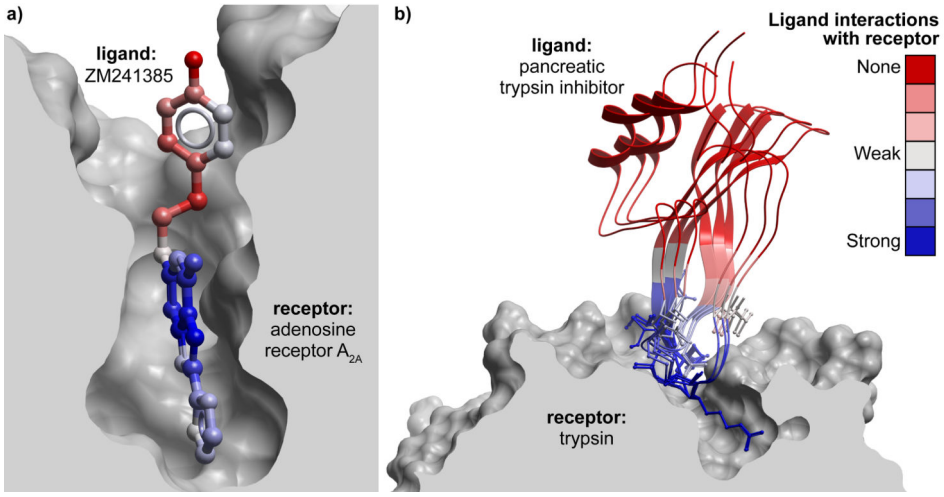

Figure 5.

Distance-based evaluation of protein-ligand (**a**) or protein-protein (**b**) complexes must be focused on ligand parts that are in direct contact with the receptor, and not on the entire ligand molecule. Because position and conformation of solvent exposed parts is only approximately defined by the interaction within the complex, such parts must be either excluded or down-weighted in docking complex evaluation.

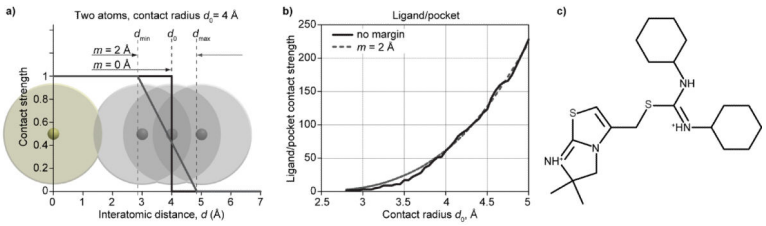

Figure 6.

Issues in evaluation of atomic contacts in protein complexes with small molecules: **(a)** definition of atomic contact strength with and without the continuous decrease margin; (**b**) hard distance cutoff (no margin) definition of the atomic contact leads to unstable behavior of the contact strength as a function of contact radius; (**c**) example of a small molecule with high degree of internal symmetry. Topologically equivalent atom permutations need to be enumerated when evaluating RMSD or comparing contacts of this molecule with its copy in a different structure.

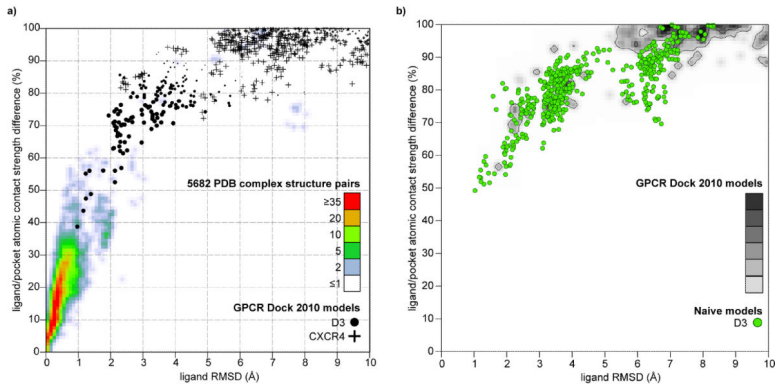

Figure 7.

Distribution of ligand RMSD values and atomic contact strength differences between identical composition complex structures: statistics of a large subset of experimental complex structures pairs in PDB (**a**, heat map), GPCR Dock 2010 models (**a**, • for D3 and + for CXCR4; **b**, heat map), and naïve models of dopamine D3 receptor (**b**, ○).

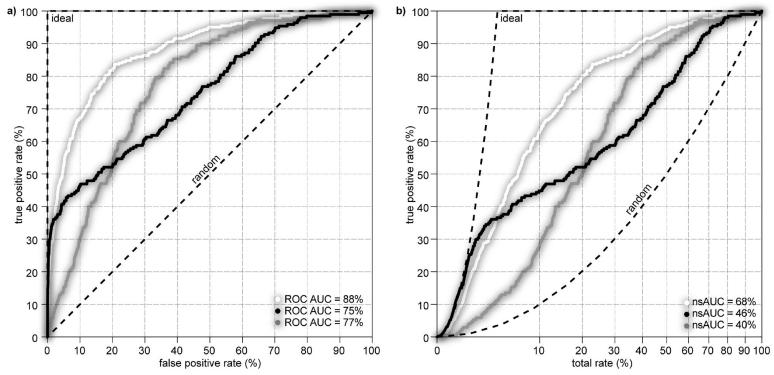

Figure 8.

Unlike the routinely used ROC AUC (**a**), the normalized square-root AUC (**b**) rewards the initial hit recognition in virtual ligand screening. This approach makes the profile in black preferable over the one in gray.

Table 1

Cumulative Distribution Function (CDF) percentiles of GPCR Dock 2010 models in the experimental distribution. Statistics are calculated for the top half of each model set, i.e. models less accurate than average are eliminated.

	Average CDF		Best CDF	
protein similarity measure	**D3**	**CXCR4**	**D3**	**CXCR4**
superimposition error	5.24%	1.68%	8.40%	2.40%
virtual C'_β-C'_β contact strength difference	2.06%	0.10%	3.99%	1.20%
ligand heavy atom RMSD	3.65%	0.91%	17.57%	5.02%
ligand-pocket contact strength difference	2.36%	0.75%	13.46%	2.60%