Chapter 3

Comparative modeling and structure prediction: application to drug discovery

Emidio Capriotti

doi:10.4155/EBO.13.192

More efficient high-throughput sequencing techniques are exponentially expanding the knowledge about the ensemble of proteins expressed by living organisms. At the same time, the determination of their 3D structure is still requiring expensive and time-consuming experiments. During the last few decades, the effort of the scientific community has allowed the crystallization of thousands of proteins, which have been resolved at the atomic level. Currently, the Protein Data Bank, the largest repository of protein structures, contains more than 88,000 macromolecular 3D structures. The computational analysis of this huge source of information revealed that during the evolution protein structure is more conserved than sequence. This finding constitutes the basic assumption behind most of the available bioinformatics algorithms for protein structure prediction. Among all the available prediction methods, those based on comparative modeling (CM) provide more accurate structures that can be used in large variety of applications, including ligand binding sites prediction and virtual screening. In this chapter, we summarize the theoretical basis and the main steps of CM. Finally, we describe their application to predict the structure of drug targets in important protein families.

Basic concepts in protein structure prediction

The classification of proteins requires three different levels of knowledge: sequence, structure and function. These three features are linked by rules that are still largely unknown. It is well known that the structure of the protein is encoded by its sequence. Indeed experimental studies have demonstrated that, after unfolding, the protein is able to assume its native 3D conformation that is responsible for the function [1]. Previous analysis of limited protein structures showed that within the same family, protein sequence is less conserved than structure [2]. The limited number of possible protein folds confirms the hypothesis that multiple proteins, generally with a common ancestor, encode for similar 3D structures. According to this observation, the solution of the protein structure prediction problem is equivalent to find the correct relationship between the space of the sequences and an exhaustive catalog of protein folds. As a consequence of this, the structure of a new protein can be predicted using the structure of a protein with similar sequence. For this purpose, it is important to define quantitative rules describing the relationship between protein sequence and 3D structure. Therefore, protein sequence alignment became a valuable method to detect evolutionary related proteins and establish empirical procedures for protein structure prediction. In general, prediction algorithms based on the detection of similarities between the unknown protein (target) and a protein with available 3D structure (template) are referred to as template based. Alternatively the template-free approaches are needed to predict the structure of new folds. Template-free methods are mainly based on physicochemical principles and information from available 3D structures [3]. Although template-free methods have broad applicability, nevertheless, their predictions are still less accurate than template-based ones. In general, template-based approaches result in high-quality models comparable with native structures. High-quality predictions from comparative modeling (CM) can be used for several applications that include the prediction of drug-binding sites and virtual screening [4–6]. This chapter focuses on CM-based structure prediction and its application on the prediction drug targeted structures.

Aa Sequence alignment: computational method that maximizes similarity between biological sequences (DNA, RNA and proteins) to detect conserved regions as possible consequence of evolutionary relationships.

Target/template: terms that indicate the protein with unknown structure (target) and the available structure (template) used as a reference in comparative modeling.

Comparative modeling: method for the prediction of protein 3D structure based on the sequence/structure similarity between target and template proteins.

Theoretical basis of CM

Protein 3D structure prediction is a hot topic in molecular biology. CM can be applied when exists a minimum level of sequence identity between the unknown protein

(target) and another protein (template) whose 3D structure is already available. CM is supported by the observations that small variations in protein sequence slightly affect protein 3D structure [2] and that accumulated mutations are constrained to conserve specific intra- and inter-molecular interactions in protein families and superfamilies [7]. The existence of highly conserved regions have been detected comparing 25 protein 3D structures from eight families [2]. The analysis of 32 pairwise alignments between homolog proteins revealed that for regions with sequence identity higher than 50%, more than 90% of Cα atoms can be superimposed with a root mean square deviation (RMSD) of approximately 1 Å, while for regions with approximately 20% sequence identity less than 42% of the structure can be superimposed with an RMSD of approximately 3 Å [2]. In the same work, it was estimated the expected rate of successfully predicted residues as a function of the sequence identity between target and template. When larger number of 3D structures became available, a more exhaustive study of the relationship between sequence and structure has been performed [8]. At the end of the 1990s, the 'twilight zone' had been defined using 792 pairwise alignments between proteins with sequence identity lower than 25% **(Figure 3.1)**. This corresponds to the low-identity region within which the sequence alignments between homolog proteins are similar to those between nonhomolog proteins. The curve separating the 'twilight zone' from the region of confident similarity has been estimated maximizing the separation between the alignments of true homologs from structurally related proteins and those of nonhomolog proteins. According to this classification, CM can be generally applied when the alignment between target and template falls in the region of confident similarity detection. This implies that for targets with no template in the confident region, the 'twilight zone' represents a limit to the application of CM.

The application of comparative modeling is supported by the observations that small variations in protein sequence slightly affect 3D structure, therefore selected mutants conserve protein structure and function.

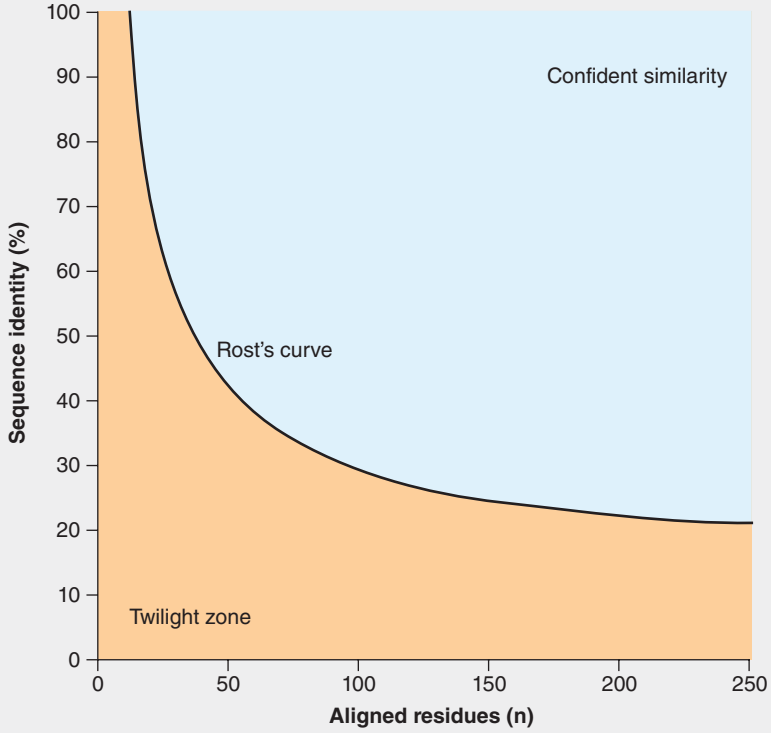
The application of comparative modeling to drug design and virtual screening is more accurate when multiple templates are available. In general, template structures representing active and inactive conformations of the protein are important to evaluate the plasticity of the target binding site.

CM protocol

CM allows the prediction of the structure of the target protein using the structure of a protein template that has a detectable level of similarity between their sequences. Accordingly, CM protocol can be summarized into four main steps **(Figure 3.2)**: selection of the template

Aa Twilight zone: region in the space of protein sequence similarity where standard alignment methods have higher failure rate in the detection of residue correspondences between target and template limiting the use of comparative modeling.

Figure 3.1. Twilight zone curve.



Reproduced with permission from [8].

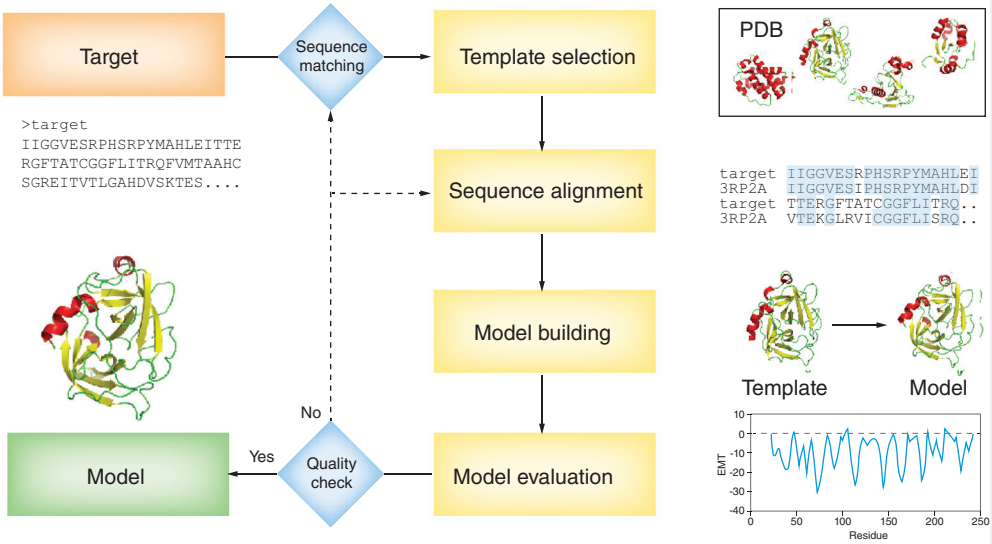
structure; sequence alignment; model building and refinement; and evaluation of the predicted structures.

Selection of the template structure

This step consists of the comparison between the target protein and a set of proteins with known structural features, searching for homologous proteins that are likely to have a similar structure. Template protein structures for this step are available at the Protein Data Bank [9], but faster searches can be performed on a reduced set from the Structural Classification of Proteins [10] and CATH [11] databases. The basic searching methods consist in pairwise alignments between target and template using BLAST (basic local alignment search tool) [12]. A further improvement of this search method is

The selection of best template structure is a key step in comparative modeling. This task is highly inaccurate in the 'twilight zone' where standard alignment methods are not able to detect similarities between target and template proteins.

Figure 3.2. Comparative modeling methods.



Reproduced with permission from [5].

the PSI-BLAST (position-specific iterative BLAST) algorithm that allows the detection of remote homolog proteins using iterative BLAST search. Recently developed methods implement profile-based algorithms, which include information from related proteins [13]. Among them, those using hidden Markov models, such as HHPred, are more accurate [14]. An overview about the available methods for the search of remote homologs has been previously published [15]. When multiple templates are available, the one with highest similarity score to the target is generally selected. Exceptions are possible when the aim of the predicted model is the study of interactions between protein and small ligand or the structure of active sites. In those cases, the templates including ligands and high-resolution structures are preferable. Therefore, the template selection is driven by considerations related to the problem the model has been built for.

Sequence alignment

The alignment between protein target and temple is a critical step for establishing the correspondences between target and template residues. In general, sequence alignment methods implement dynamic-programming algorithms that use the BLOSUM (blocks substitution matrix) [16] and the

PAM (point accepted mutation) [17] substitution scoring matrices. For proteins with high level of similarity, sequence alignment methods tend to return similar results. If the sequence identity drops down 40%, more accurate alignments should include structural information and multiple sequence alignments of homolog proteins. In this case the alignments obtained through automatic methods need to be manually checked.

Model building & refinement

In this step the 3D structure of the target protein is predicted using the correspondences between aligned residues obtained in the previous step. CM algorithms can be grouped in three classes: segment matching, rigid body assembly and spatial restraints satisfaction. These classes differ in the method used to transfer structure information from the template to the target. Rigid body assembling and segment matching use coordinates and conformations from conserved regions or matching peptides in the template structure. Methods based on spatial restraints transfer atomic restraints from the template protein to the equivalent atoms in the target protein, including a procedure that optimizes the search of the low-energy conformations minimizing the number of violated restraints. The predictions of loop and side-chain conformations represent the most difficult tasks. In particular, the structural variability of loop regions is caused by frequent residue insertions and deletions. Thus, specific methods have been implemented to predict loop and side-chain conformations. In the final step, the predicted structure is refined optimizing the conformations of the residues at the interface between nonconserved and conserved regions. This task can be performed by molecular dynamic (MD) simulations, which use an interatomic force field to improve the quality of predicted models.

Evaluation of the predicted structures

The evaluation of predicted 3D structures obtained by CM protocol consists of two steps: evaluation of geometry and the stereochemistry of the predicted model, and evaluation by statistical potentials [18]. The geometry of the predicted models is analyzed to check if bond distances and angles are correct and to avoid steric clashes. Methods based on statistical potentials evaluate the interactions of each atom in the model and compares them with the average atomic interactions in high-resolution structures. Although the theoretical bases of statistical potentials are still questioned, they are currently used for the model assessment and selection of high-quality predictions. Similar methods use standard MD simulation

The quality of predicted structures is strongly dependent on the level of sequence similarity. Higher sequence similarity between target and template proteins generally results in more accurate models.

force fields to evaluate the quality of predicted structures. Depending on the results of the evaluation, it is possible to repeat the first two steps selecting a better template or improving the sequence alignment. Thus, the prediction process can be iterated until the model obtains the best results in the evaluation step.

Examples of extensively used and freely available tools for CM are I-TASSER, MODELLER and Robetta. A selected list of available resources and methods and resources for CM are reported in **Table 3.1**.

CM for drug design

The knowledge of protein 3D structure information is key in drug design enabling the selection of a subset of ligands, which can potentially bind a given target. This procedure, referred to as virtual screening, is extensively adopted to reduce the cost of time-consuming and expensive assays for the design and repurposing of new therapeutics. The virtual screening procedure consists of the determination of binding-site residues where the ligand is docked and scored to estimate the binding affinity. The relative orientation between target and ligand is predicted by rigid-body or flexible docking of their 3D structures. The increasing computational power is making flexible docking more affordable allowing to sample different ligand-target conformations. In a recent work [19], a set of models for 21 x-ray protein–ligand complexes in CCDC/Astex test set [20] has been selected to estimate the expected quality of docking complexes obtained using predicted structures by CM. The results reveal that models with sequence identity higher than 50% show a RMSD value lower than 2 Å with respect to experimental x-ray structures. In addition, for a large fraction of these models the local RMSD for the binding site atoms is also lower than 2 Å. These results confirm that state-of-the-art methods for structure prediction are effective tools for modeling the interactions between ligand and protein target.

In the following sections, the interesting cases of G-protein-coupled receptors (GPCRs) and protein kinases target families are discussed.

G-protein-coupled receptors

The GPCRs constitute the most abundant protein superfamily among transmembrane proteins. Sequence analysis algorithms revealed that approximately 800 human genes encode for proteins belonging to the GPCR superfamily (~4% of the human protein-coding genome). A classification scheme for GPCRs divided them in six main classes with low level of sequence similarity. Approximately 85% of GPCR genes encode protein in class A, also referred to as the rhodopsin family. Since GPCRs represent a large target family, accounting for 20–50% of approved drugs, the knowledge of their 3D

Table 3.1. Computational methods and resources for protein structure prediction.

Name	URL
Repositories and resources for comparative modeling	
ModBase	http://modbase.compbio.ucsf.edu
Protein Model Portal	www.proteinmodelportal.org
SWISS-MODEL Repository	http://swissmodel.expasy.org/repository
Resources for GPCRs and protein kinases	
GPCR Database	www.gpcr.org/7tm
GPCR Research Database	http://zhanglab.ccmb.med.umich.edu/GPCRRD
Kinome^{LHM}	http://cssb2.biology.gatech.edu/kinomelhm
Protein Kinase Resource	http://pkr.genomics.purdue.edu
Structure and classification databases	
CATH	www.cathdb.info
Protein Data Bank	www.pdb.org
Pfam	http://pfam.sanger.ac.uk
Structural Classification of Proteins	http://scop.mrc-lmb.cam.ac.uk/scop
Template selection	
Basic Local Alignment Search Tool	http://blast.ncbi.nlm.nih.gov/Blast.cgi
FASTA	www.ebi.ac.uk/Tools/fasta
HHPred	http://toolkit.tuebingen.mpg.de/hhpred
SAM-T08	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html
Threader	http://bioinf.cs.ucl.ac.uk/threader
Sequence alignment methods	
CLUSTALW	www.ebi.ac.uk/Tools/msa/clustalw2
MAFFT	http://mafft.cbrc.jp/alignment/server
MUSCLE	www.drive5.com/muscle
T-Coffee	www.tcoffee.org

GPCR: G-protein-coupled receptor.

Table 3.1. Computational methods and resources for protein structure prediction.

Name	URL
Tools for comparative modeling	
I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER
Modeller	www.salilab.org/modeller
ModWeb	https://modbase.compbio.ucsf.edu/scgi/modweb.cgi
Robetta	http://robetta.bakerlab.org
SWISS-MODEL	http://swissmodel.expasy.org
Methods for model evaluation	
ANOLEA	http://melolab.org/anolea
DFIRE	http://sparks.informatics.iupui.edu/yueyang/DFIRE
PROCHECK	www.ebi.ac.uk/thornton-srv/software/PROCHECK
ProSa-web	https://prosa.services.came.sbg.ac.at
QMEAN	http://swissmodel.expasy.org/qmean

GPCR: G-protein-coupled receptor.

structure is extremely important for designing new drugs. Although recent advances have been made in the crystallization of new GPCRs the structural characterization of the whole superfamily is still incomplete. Therefore, the prediction of unknown GPCRs by CM is essential for the screening of new drugs. Since GPCRs share low level of sequence similarity the key step in CM is the selection of the best template. Currently high-resolution crystallographic data are available for 11 class A proteins (see **Table 3.2**) and one class B GPCR. The consistency of available templates makes CM suitable only for class A GPCRs. The bovine rhodopsin is the most studied structure for GPCR but unfortunately it is distant in sequence homology to other class A GPCRs. Thus, the use of rhodopsin x-ray structure as templates for CM can result in errors in the sequence alignment. Another challenging task in CM consists of the accurate prediction of the binding sites that can adopt different conformations depending on the function of the ligand. Recent studies of the binding regions of rhodopsin [21] and the β2-adrenergic receptor [22] provide important insight about the conformational changes related to their activation. Available templates for active and inactive states facilitate the application of CM to other GPCRs showing similar interactions. In contrast to these limitations in the prediction of GPCR structures using rhodopsin templates, successful examples proved that available biochemical insights improve the accuracy of predicted models. This type of information can be included as spatial restraints during the modeling procedure. The resolution

Table 3.2. Class A G-protein-coupled receptor structures in the Protein Data Bank.

Protein name	Protein Data Bank code
Rhodopsin	1F88, 1HZX, 1L9H, 1GZM, 1U19, 2HPY, 2G87, 2I35, 2I36, 2I37, 2J4Y, 2PED, 2ZIY, 2Z73, 3CAP, 3C9L, 3C9M, 3DQB, 3PXO, 3PQR, 2X72
Adenosine-A_{2A} receptor	3QAK, 3EML, 2YDO, 2YDV
β1 adrenergic receptor	2VT4, 2Y00, 2Y02, 2Y03, 2Y04, 2Y01, 2YCW, 2YCX, 2YCZ, 2YCY
β2 adrenergic receptor	2RH1, 2R4R, 2R4S, 3D4S, 3NYA, 3NY8, 3NY9, 3PDS, 3P0G, 3SN6
CXCR4 chemokine receptor	3OE0, 3OE6, 3OE8, 3OE9, 3ODU
Dopamine receptor 3	3PBL
Histamine receptor 1	3RZE
M3 muscarinic acetylcholine receptor	4DAJ
Kappa opioid receptor	4DJH
Nociceptin/orphanin FQ receptor	4EA3
Sphingosine 1-phosphate receptor	3V2W

of new GPCR structures and the characterization of their alternative conformations have been crucial for the understanding of the relationship between sequence and structure in the presence of different ligands. For example, the structure of CXCR4 adrenergic receptor showed a larger and more open binding site closer to the extracellular surface when compared with β2-adrenergic receptor and rhodopsin. Such differences make CXCR4's binding region able to bound different ligands, suggesting a degree of variability in the local structures of GPCR binding regions. The systematic analysis of known GPCR structures indicated that they only represent a fraction of all the conformations assumed by class A GPCRs. Thus, the structural variability of the GPCRs suggests that more accurate predictions can be obtained using multiple templates. In addition, MD can be useful to sample alternative structural conformations and improve model refinement.

An interesting online resource for GPCRs structure prediction is the GPCRRD database, which collects experimental restraints from the literature. In the near future, it is expected that the increasing number of experimental data and available template structures will results in advancements in GPCR CM.

Protein kinases

The protein kinases constitute a large family of enzymes, accounting for approximately 2% of the human proteome. These proteins are involved

in many cellular processes such as inflammation, differentiation, proliferation and apoptosis, and therefore they are targets of several therapeutic strategies. Data collected at the Protein Kinase Resource includes more than 450 3D structures, approximately 65% of which are humans. A recent estimation revealed that more than 500 different human protein kinases exists. A classification of kinases according to the sequence similarity of their catalytic domains grouped them into eight major kinase families and 'others' or 'atypical' groups, including all the remaining ones [23]. An alternative scheme based on substrate preferences divides protein kinases into serine/threonine, tyrosine, histidine and aspartic/glutamic kinases. The level of sequence/structural identities within the kinase families makes unsolved proteins ideal candidates for CM and for drug design. The activation state of the protein kinases is determined by the conformation assumed by activation loop. The two alternative states are characterized by different structural rearrangements of the catalytic site. Although many studies focused on the characterization of the active conformation, many inhibitors interact with the inactive forms that are highly variable across dissimilar kinases. The structural plasticity of the inactive site limits the application of CM in virtual screening because predicted structures based on the active conformation do not differ significantly from the templates. Even under this limitation, CM has been applied successfully to the prediction of protein kinases for nonvirtual and virtual screening. A recent work summarizes the results of a large-scale *in silico* screening of the whole human kinome using sequence profile alignments of ligand-free and ligand-bound conformations [24]. The computational analysis of approximately 2 million ligands resulted in the screening of approximately 5 million ligand-target complexes ranked by different scoring functions. The quality of the modeling procedure was evaluated comparing of structural predictions against the native structures of the active (holo) and the inactive (apo) forms of human kinases. The results showed an average RMSD of 2.7 Å and 3.1 Å, respectively, for the holo and apo conformations. The lower RMSD obtained for active versus inactive forms reflects the higher structural variability of templates in the holoconformations. The comparison of the kinase binding regions showed an RMSD of approximately 2 Å for the all atoms representation. This result is in agreement with the predicted plasticity of the binding site that allows members of the same kinase family to bind similar ligands.

Recent reviews describe the application of CM procedures for virtual screening [25,26]. Their effectiveness is demonstrated by successfully applications to GPCRs and protein kinases that have been reported [27,28].

Conclusion & future perspective

Protein structure prediction by CM is largely used in many practical tasks. During the last years, the continuous improvement in overall accuracy of the predicted models has made virtual screening and drug design procedures more effective. The exponential increase of protein sequences from high-throughput technologies results in a higher number of predicted models that need to be evaluated with fast and accurate tools. In addition, the large amount of data generated by more powerful computational devices enables to perform exhaustive search in conformational space of target-ligand complexes. Therefore, it will be important to develop highly curated databases collecting both experimental and *in silico* data. In this direction is the ChEMBL database [101], which integrates chemical and genetic information for GPCRs and protein kinases. In the near future, it is expected that well-curated and integrated structure data will be key for the selection of new potential targets and the development of new drugs.

Financial & competing interests disclosure

The author has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Summary.

- Comparative modeling is the most accurate method for protein structure prediction based on the sequence/structure similarities between the unknown protein (target) and a protein with known structure (template).
- The application of comparative modeling is limited by the level of sequence similarity between target and template.
- The twilight zone defines the region where the sequence/structure similarities between target and template are difficult to detect by standard alignment methods.
- Comparative modeling consists of four main steps: template selection, sequence alignment, model building and model evaluation.
- The quality of the predicted structure (model) strongly depends on the sequence similarity between target and template.
- The selection of a good template is driven by considerations related to the resolution of the problem for which the model has been built for.

References

1 Anfinsen CB. Principles that govern the folding of protein chains. *Science* 181(4096), 223–230 (1973).

2 Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5(4), 823–826 (1986).

3 Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.* 17(3), 342–346 (2007).

4 Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 294(5540), 93–96 (2001).

5 Liu T, Tang GW, Capriotti E. Comparative modeling: the state of the art and protein drug target structure prediction. *Comb. Chem. High Throughput Screen.* 14, 532–537 (2011).

6 Lahti JL, Tang GW, Capriotti E, Liu T, Altman RB. Bioinformatics and variability in drug response: a protein structural perspective. *J. R. Soc. Interface* 9(72), 1409–1437 (2012).

7 Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* 10(10), 709–720 (2009).

8 Rost B. Twilight zone of protein sequence alignments. *Protein Eng.* 12(2), 85–94 (1999).

9 Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* 35(Database issue), D301–D303 (2007).

10 Andreeva A, Howorth D, Chandonia JM *et al.* Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 36(Database issue), D419–D425 (2008).

11 Cuff AL, Sillitoe I, Lewis T *et al.* The CATH classification revisited – architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.* 37(Database issue), D310–D314 (2009).

12 Altschul SF, Madden TL, Schaffer AA *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17), 3389–3402 (1997).

13 Capriotti E, Fariselli P, Rossi I, Casadio R. A Shannon entropy-based filter detects high-quality profile–profile alignments in searches for remote homologues. *Proteins* 54(2), 351–360 (2004).

14 Soding J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* 21(7), 951–960 (2005).

15 Fariselli P, Rossi I, Capriotti E, Casadio R. The WWWH of remote homolog detection: the state of the art. *Brief. Bioinform.* 8(2), 78–87 (2007).

16 Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* 89(22), 10915–10919 (1992).

17 Dayhoff MO, Schwartz R, Orcutt BC. A model of evolutionary change in proteins. In: *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington DC, USA, 345–358 (1978).

18 Capriotti E, Marti-Renom MA. Assessment of protein structure predictions. In: *Computational Structural Biology: Methods and Applications.* Schwede T, Peitsch MC (Eds). World Scientific Publishing Company, Singapore, 89–109 (2008).

19 Bordogna A, Pandini A, Bonati L. Predicting the accuracy of protein–ligand docking on homology models. *J. Comput. Chem.* 32(1), 81–98 (2011).

20 Nissink JW, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein–ligand interaction. *Proteins* 49(4), 457–471 (2002).

21 Standfuss J, Edwards PC, D'Antona A *et al.* The structural basis of agonist-induced activation in constitutively active rhodopsin. *Nature* 471(7340), 656–660 (2011).

22 Rasmussen SG, Choi HJ, Fung JJ *et al.* Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. *Nature* 469(7329), 175–180 (2011).

23 Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. *Science* 298(5600), 1912–1934 (2002).

24 Brylinski M, Skolnick J. Comprehensive structural

and functional characterization of the human kinome by protein structure modeling and ligand virtual screening. *J. Chem. Inf. Model.* 50(10), 1839–1854 (2010).

25 Rockey WM, Elcock AH. Structure selection for protein kinase docking and virtual screening: homology models or crystal structures? *Curr. Protein Pept. Sci.* 7(5), 437–457 (2006).

26 Yarnitzky T, Levit A, Niv MY. Homology modeling of G-protein-coupled receptors with x-ray structures on the rise. *Curr. Opin. Drug Discov. Devel*. 13(3), 317–325 (2010).

27 Carlsson J, Coleman RG, Setola V *et al.* Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nat. Chem. Biol.* 7(11), 769–778 (2011).

28 Sandberg EM, Ma X, He K, Frank SJ, Ostrov DA, Sayeski PP. Identification of 1,2,3,4,5,6-hexabromocyclohexane as a small molecule inhibitor of JAK2 tyrosine kinase autophosphorylation [correction of autophophorylation]. *J. Med. Chem.* 48(7), 2526–2533 (2005).

Website

101 ChEMBL Database. www.ebi.ac.uk/chembldb