# *Computational methods for the elucidation of protein structure and interactions*

# Computational Methods for the Elucidation of Protein Structure and Interactions

Nicholas S. Edmunds [1]

Liam J. McGuffin [1]✉

Email: l.j.mcguffin@reading.ac.uk

[1] School of Biological Sciences, University of Reading, Reading, UK

## Abstract

Biologists are increasingly aware of the importance of protein structure in revealing function. The computational tools now exist which allow researchers to model unknown proteins simply on the basis of their primary sequence. However, for the non-specialist bioinformatician, there is a dazzling array of terminology, acronyms, and competing computer software available for this process. This review is intended to highlight the key stages of computational protein structure prediction, as well as explain the reasons behind some of the procedures and list some established workarounds for common pitfalls. Thereafter follows a review of five one-stop servers for start-to-finish structure prediction.

AQ1

## Key words

Tertiary structure
Homology modeling
Template-based
Template-free
Sequence
Alignment
Refinement
Quality assessment
Docking
Quaternary structure

# 1. Introduction

Understanding macromolecular 3-D structure remains a major ambition for molecular biologists. This is due, not only to the therapeutic potential offered by nucleic acid–protein and protein–ligand interactions as new medicinal drug targets, but also to many wider applications of protein structure knowledge including agricultural crop improvement or even biofuel development [1].

Computational or in silico methods for the determination of protein structure are becoming ever more widespread and important in fulfilling this ambition. This is fundamentally the consequence of two phenomena: firstly, that the ability to elucidate protein sequences from genomic information continues to outpace the capability of experimental methods to determine the structure of these newly sequenced proteins [1], despite advances in X-ray crystallography technique and improvements of NMR and cryo-EM accuracy and resolution; and secondly, the continuing assertion that structure implies function in protein biology and that, in turn, sequence determines structure. Therefore, the sequence to structure gap continues to grow and manual experimental techniques are unlikely to close this in the near future [2].

AQ2

AQ3

Since the creation of the Protein Data Bank (PDB) [3] in 1971, there has been an increasing reliance on curated sequence and structural repositories by the molecular biology community. Furthermore, along with community-wide experiments such as CASP (Critical Assessment of techniques for protein Structure Prediction) and CAPRI (Critical Assessment of the PRediction of Interactions)—see Subheading 6 for more details, growth in the area of in silico methods has led to an explosion in predicted protein structures [4]. This has mainly occurred through the rise of homology (or template-based) modeling and has in turn driven the associated proliferation of prediction software and data repositories, which are now available to research communities via the internet.

In this chapter, we will attempt to explain some of the main techniques used in 3-D protein structure prediction along with decoding a number of acronyms commonly encountered within the field; and secondly, to clarify the wide array of software packages and databases that now exist and, in the process, reference and analyze some key representative examples.

## 2. A Brief Summary of Protein Classification and Data Repositories

Proteins can be classified in a number of ways; in terms of primary structure or sequence similarity; secondary structure and associated motifs; tertiary structure and associated folds and domains and an emerging categorization based on protein–protein interactions (PPI) [1]. In addition, and perhaps related more closely to secondary structure classification than any of the others, is the grouping of proteins into classes and families on the basis of evolutionary relationship. The following describes a little about resources that fall into these classification categories.

In the case of primary structure, there are a number of databases containing information on amino acid sequences of which probably the most important from a structural prediction point of view is the Protein Knowledge Base—UniProtKB/TrEMBL [5]. This vast protein sequence database consists of the Universal Protein Resource (UniProt from PIR) which evolved from the early manually annotated SWISS-Prot sequence database (1986) allied to the automatically annotated TrEMBL sequence database administered by the European Bioinformatics Institute (EBI). The resource also contains UniRef a clustering service which lists groups of related sequences together and UniParc, an additional development intended to represent a complete and comprehensive non-redundant database of all known protein sequences with each sequence listed only once with a unique identifier (see Table 1). Tools for assessing sequence similarity and alignment based on sequence database searches are discussed in Subheading 4.1 below.

**Table 1**

Protein sequence databases

AQ4

| Name | Description | Website |
|---|---|---|
| UniProtKB [5] | Repository for sequence, taxonomy, annotation, ontology, and classification information including TrEMBL (automatically annotated sequences) | www.uniprot.org/help/uniprotkb |
| UniParc [5] | Non-redundant database of all known protein sequences | www.uniprot.org/help/uniparc |
| UniRef [5] | Clustering service of related sequences | www.uniprot.org/help/uniref |

Information on classifying proteins according to secondary structure is most easily obtained from the structural classification repositories [1]; Pfam [6] (from the EBI, classifies proteins into families based on domain similarity), SCOP [7] (Structural Classification Of Proteins—classifies into family, superfamily, and fold similarity), and CATH [8] (from UCL, classifies proteins into class, architecture, topology, and homologous families on the basis of domain similarity) and each of these has a website with full information on their classification system and how best to interpret it. These databases contain a great deal of evolutionary and relationship information as well as links to other software and are widely referenced by many 3-D prediction algorithms.

For novel protein sequences whose structures are not recorded in any existing database, the most widely accepted methods of secondary structure prediction (also referenced below) are those based on the Dictionary of Protein Secondary Structure algorithm (DSSP) [9] and these include PSIPred [10] and JPred4 [11] although it is possible to find many others via links within the ExPASy Bioinformatics Resource Portal.

The major resource for known tertiary structure information is, of course, the PDB (Protein Data Bank) [12] although a number of alternative databases can be found including those at the NCBI and EBI webpages (see Table 2). These have links to many classification and prediction resources. Again, the SIB (Swiss Institute of Bioinformatics) resource portal ExPASy may be useful with links to nextProt [14] (a human protein knowledge base), STRING [15] as well as Swiss-Model [16] (see Subheading ~~6.5~~ 5.5).

**Table 2**

Protein structure and classification databases

| | | |
|---|---|---|
| CATH [8] | Structural classification into class, architecture, topology, and homology | www.cathdb.info/ |
| Pfam [6] | Protein family classification (EBI) | https://pfam.xfam.org/ |
| SCOP [7] | Structural classification of proteins | http://scop.mrc-lmb.cam.ac.uk/scop/ |
| PDB [3] | The protein data bank, from wwPDB, a collaboration of PDBe (UK), PDBj (Jpn), and BMRB (US) | www.rcsb.org/pdb |
| PDBe-PISA [13] | Proteins, interfaces, structures, and assemblies database for protein–protein interactions and quaternary structures | www.ebi.ac.uk/msd-srv/prot_int/pistart.html |

| | | |
|---|---|---|
| nextProt [14] | Human protein knowledge base | https://www.nextprot.org/ |
| STRING [15] | Alternative protein–protein interaction knowledge base from the SIB | https://string-db.org/ |

AQ5

Probably the most comprehensive quaternary and protein–protein interaction database is PDBe-PISA [13] (Proteins Interfaces, Structures, and Assemblies) that is hosted by the EBI although SIB's SMTL (Swiss-Model Template Library) [16] and STRING are also useful for studying interactions and networks.

## 3. Types of Structure Prediction; Comparative Versus Ab Initio Modeling

The most successful form of structure prediction to emerge over the last 25 years is comparative modeling [12]. At its most basic, this is the process of modeling a protein with an unknown tertiary structure on the basis of sequence similarity to those with known structures.

Proteins that have a matching sequence (sequence identity above 30% as a rule of thumb) [17] are deemed homologs and can be used as templates on the presumption that sequence similarity suggests a common functional evolutionary ancestor. A similar structure can therefore be inferred from a similar sequence.

This approach is known variously and almost interchangeably as Comparative Modeling (CM), Homology Modeling (HM), and Template-Based Modeling (TBM) (although true homology modeling relies on an established evolutionary relationship between proteins rather than just a distant sequence similarity or shared domain). For the rest of the chapter, we will refer to this process as Template-Based Modeling or TBM.

Ab initio modeling, on the other hand attempts to use the so-called physics-based rules and routine, e.g., torsion angles in the protein carbon backbone, hydrophobicity ratings, bond length calculations, and van der Waals interactions, to predict the folding and hence tertiary structure of a protein from sequence alone, i.e., without comparison with a template [18]. This is often alternatively termed de novo modeling, although, strictly speaking de novo modeling may include some type of sequence fragment check against a database whereas true ab initio techniques should model from sequence alone. A complication that might be

encountered is that a number of programs now include a certain level of ab initio modeling embedded within their TBM calculations (e.g., the Rosetta algorithm [2, 19]) or to help resolve unstructured parts of the suggested model (e.g., Phyre2 [20]). However, there are other programs that offer a complete ab initio modelling service (e.g., QUARK, FALCON as well as ROSETTA).

The following sections will concentrate on describing TBM only, as this is likely to be the most useful route for the general molecular biologist who is not part of a specialist protein modeling group, and the technique is applicable to the majority of new protein targets.

## 4. Stages in Template-based Modeling (TBM)

TBM is a multi-step process [1], often made to appear seamless by publicly accessible webserver programs (see Table 3 below for a list). However, the identification of suitable homologs to use as templates is often not an insignificant task, and there are a number of technical solutions employed across various platforms to ensure that the templates used in model building are as relevant as possible. Another problematic stage in the modeling process is the sorting, scoring, and ranking of the (often) many alternative models (termed decoys) that are built [2]. These two stages remain the greatest challenge in TBM with the latter potentially more challenging than the former due to the nature of selecting the closest model to the native protein whose structure is unknown.

**Table 3**

Tertiary structure prediction tools

| | | |
|---|---|---|
| IntFOLD [21] | A high-performance server developed by the McGuffin group, offering a suite of programs for tertiary and quaternary predictionSpecializing in model quality assessment | https://www.reading.ac.uk/bioinf/IntFOLD/ |
| I-TASSER [22] | A powerful threading-based online server offering a number of services in addition to modeling | https://zhanglab.ccmb.med.umich.edu/I-TASSER/ |
| MODELLER [23] | Downloadable program for 3-D structure prediction. Users must provide their own alignment data | https://salilab.org/modeller/ |
| MULTICOM [24] | Part of an online toolbox for structure prediction hosted by | http://sysbio.rnet.missouri.edu/multicom_cluster/ |

| | | |
|---|---|---|
| | the university of Missouri | |
| Pcons [25] | Online server specializing in quality assessment (Stockholm university) | http://pcons.net/ |
| Phyre2 [20] | Online full-service server from the structural bioinformatics Group at Imperial College part of Genome3D | http://www.sbg.bio.ic.ac.uk/phyre2 |
| Predict protein (PP) [26] | Developed by RostLab (university of Munich) offering full prediction service | https://www.predictprotein.org/ |
| RaptorX [27] | Online server (Xu group, University of Chicago), specializing in predicting sequences with no close homologs | http://raptorx.uchicago.edu/ |
| ROBETTA [28] (Rosetta [19, 29]) | Online server (Baker lab, University of Washington) full structure prediction using the powerful Rosetta algorithm | http://robetta.bakerlab.org/ |
| SWISS-MODEL [16] | Comprehensive online server; both tertiary and protein interaction prediction by the SIB (Swiss Institute of Bioinformatics) | https://swissmodel.expasy.org/ |

Rangwala and Kapris, 2010 [1] split the process of TBM (comparative modeling in their review) into five distinct stages: Selection of templates, Alignment of sequences, Model building, Quality evaluation, and Refinement, and in the following sections we have highlighted a similar but updated sequence of events routinely used by the protein modeling community.

The flowchart below gives an overall guide to the way the sequence fits together and the decision points that drive the process. It must be noted, however, that these stages are in-built and often invisibly merged in most public webservers making it unclear which distinct stage is being carried out at any one time. For those wishing to perform TBM in a more hands-on manner, there are specialist programs which can be downloaded and run separately from many of the website listed in Table 3, but for most non-specialist bioinformaticians these sections represent background information as the majority of your modeling needs will be catered for by using the full structure prediction webservers described in Subheading 5.

## 4.1. Sequence Alignment and Template Identification

The initial task is that of identifying one or more suitable homologs to use as templates on which to base the model (see Table 4 for a list of programs). The amino acid sequence of the protein of interest, the target protein, will be run against a database of sequences, often the UniprotKB or a non-redundant derivative thereof. Here, the first problem is encountered; evolutionarily related proteins often have a greater level of structure conservation than sequence conservation [20]. Therefore, it is possible that simply aligning the whole of your target sequence against a sequence from another protein will produce a poor match. Most sequence alignment programs (e.g., Uniprot-align [5] and PSI-BLAST [30]) will therefore attempt local sequence alignment where sequences are cut into sections that are then cross-aligned [35]. The rationale is that protein domains may swap places over time and therefore one needs to search the whole sequence for matches rather than a simple pairwise comparison. Even with successful alignments there is a high probability of missing sequence sections (deletions), additional sections (insertions), and substitutions where amino acids have been replaced with others. For this reason, sequence alignments are scored from a BLOSUM matrix [18] that attempts to give good scores for amino acid conservation or replacement in non-structured parts of the protein (loop regions) and penalties for missing sections or replacement of amino acids in ordered secondary structure regions. A number of programs will also employ a secondary structure consensus check between target and templates at this stage [20] to increase confidence in final template selection, a popular choice of program being PSIPRED (UCL).

**Table 4**

Protein sequence search and alignment tools

| | | |
|---|---|---|
| BLAST [30] | Basic local alignment tool (also see PSI-BLAST a more sensitive version) | https://blast.ncbi.nlm.nih.gov/ |
| ClustalW [31] | Multiple sequence alignment using traditional sequence profiling | https://embnet.vital-it.ch/software/ClustalW.html |
| Clustal Omega [32] | Multiple sequence alignment tool using HMM profiling | https://www.ebi.ac.uk/Tools/msa/clustalo/ |
| EMBOSS [33] | Global alignment (needle option) and local alignment (water option) | https://www.ebi.ac.uk/Tools/psa/emboss_needle/emboss_water/ |
| FASTA [33] | A simple local alignment tool | https://www.ebi.ac.uk/Tools/sss/fasta/ |
| HH-blits [34] | Popular hidden Markov model (HMM) alignment site | https://toolkit.tuebingen.mpg.de/tools/hhblits |
| HMMER | Sequence search tool using hidden | http://hmmer.org/ (to download) |

| | | |
|---|---|---|
| [34] | Markov models (HMM) prediction | https://www.ebi.ac.uk/tools/hmmer/search/phmmer (online) |

## 4.2. Loop Identification and Side-Chain Packing

Many homologous proteins will share not only a certain agreement in sequence identity but also in secondary structure, folds, and overall configuration. However, it is quite frequent for related proteins to differ in the length of the unstructured loop regions that connect secondary structure as well as the order of the individual folds or domains. For this reason, researchers have often been obliged to take the extra step of loop building in order to account for longer or shorter unstructured regions between folds. Many contemporary programs now include loop building as an automatic function [20], but optimization of loops and unstructured regions still occurs in refinement programs (see below). Side-chain packing is another element of model building which has become absorbed into the regular functioning of modern modeling programs [36], but which is still an important part of refinement procedures. Often the last part of refining a model will be to assess clashes or unlikely contacts between amino acid side chains and attempt to modify angles and residue positions slightly in order to resolve these.

## 4.3. QA and Ranking Models

Once models are constructed by the modeling software the importance of assessing their quality is necessary for two reasons. The first, which is discussed further in the following section, is to rate the models on general agreement with known protein structures, in other words, have you built a native-like potentially functional model or is it so far beyond acceptable structural limits as to be unlikely to exist? The second is the task of assessing which of your models matches your protein's native structure the best and therefore should be at the top of your ranking list.

In general, single-model quality assessment methods (those assessing each model individually) employ a number of physical checks to assess the models' structural integrity. These range from residue environment compatibility, e.g., hydrophobicity and solvent accessibility to structural features, such as secondary structure compatibility and assessment of backbone torsion angles [12]. Users are then presented with scores showing how well the model conforms to hypothetical 3D norms. One problem that must be borne in mind when interpreting these plausibility checks is that a model may score well because it conforms to pre-programmed

ideals and so be ranked above a model which displays some structural defects but nevertheless is much closer to the native structure.

The second issue of ranking models may be relatively simple if all that was required was to select the best model on the basis of its resemblance to the template. However, with lower sequence identities the key question becomes, how closely does resemblance to the template suggest closeness to the native structure? Ranking models' resemblance to a native structure that is unknown will always be a subjective process and so consensus assessment has been developed in an attempt to overcome this.

Consensus methods use scores from a number of different programs, and many include a clustering stage in which models are clustered together on the basis of structural similarity, selecting those that lie close to the largest clusters. Consensus assessment can often out-perform single methods, with clustering working well when templates and models show a close structural relationship [37]. However, if there is a large variability in templates leading to a significant number of low-quality models or very few models in the first place, clustering and consensus methods that include them can prove less reliable.

As can be imagined, the distinction between the disciplines of assessment for ranking and final model quality assessment has become blurred and the processes now overlap somewhat.

Model quality is, to a large extent, dependent on the evolutionary distance between the target protein and the template(s) used to model it [1]. When working with low sequence identity, target-template 3-D similarity naturally decreases meaning that models may contain significant errors. As stated, model quality assessment assigns a predictive score to a model [12] in an attempt to rate its accuracy or similarity to the native protein prior to any confirmatory experimental structure being available and over the years a number of approaches have been developed.

Early versions of quality checks focused on stereochemical calculations measuring, amongst others, bond angles, steric clashes, and Ramachandran outliers. Others were based on calculating an energy score based on the model's perceived distance from a hypothetical free energy minimum. The so-called energy function checks fell broadly into two groups: those calculating a statistical score by analyzing the model against known protein structures and those calculating an empirically derived energy score from force field and molecular dynamic data. The shortcomings of these quality checks were, as mentioned before, that models could

have perfectly reasonable stereochemical profiles and a low energy conformation but neither guaranteed similarity to the unknown native structure.

Current MQAPs (a selection listed in Table 5) attempt to overcome these shortcomings by combining a number of approaches. Firstly, as well as giving a global score for the overall model many programs will also give a local, or per residue score which assesses each amino acid residue and the favorability of the surroundings in which it finds itself in the proposed chain (factors like solvent accessibility, secondary structure compatibility, and side-chain contacts may be assessed). Secondly, in addition to basic stereochemical checks and energy considerations most MQAPs will perform a clustering routine [37] where potential models (decoys) are clustered on the basis of their conformation similarities. Models representative of large clusters are assumed to have a higher likelihood of resembling the native structure than remote models. Lastly, to increase the statistical confidence of the final score, neural networks can be used to perform an all-against-all comparison of conformations and then calculate a probability score [12]. The advantage of using neural networks is not only their ability to handle vast amounts of data but also the ability to train the networks to recognize native conformations from decoys using a training set of experimentally solved structures.

**Table 5**

A selection of Model Quality Assessment Program servers (MQAPs)

| | | |
|---|---|---|
| ModFOLD6 [21] | A resource for estimates of model accuracy (EMA), using a hybrid quasi-single model approach | https://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD6_form.html |
| PCons [25] | Analyses models for recurring 3-D structural patterns and assigns a commonality score | http://pcons.net/index.php?about=pcons |
| ProQ3 [38] | Based on Rosetta, including all-atom (ProQRosFA) and centroid (ProQRosCen) energy functions | http://proq3.bioinfo.se/ |
| QMEAN [39] | The sum of four measures; backbone torsion angles, Cβ interactions, all atom interactions, solvation score | https://swissmodel.expasy.org/qmean/ |

See ~~Appendix H~~ Section 8 Notes (Table 11) for a table of scores commonly encountered with model quality assessment, refinement, and ranking output.

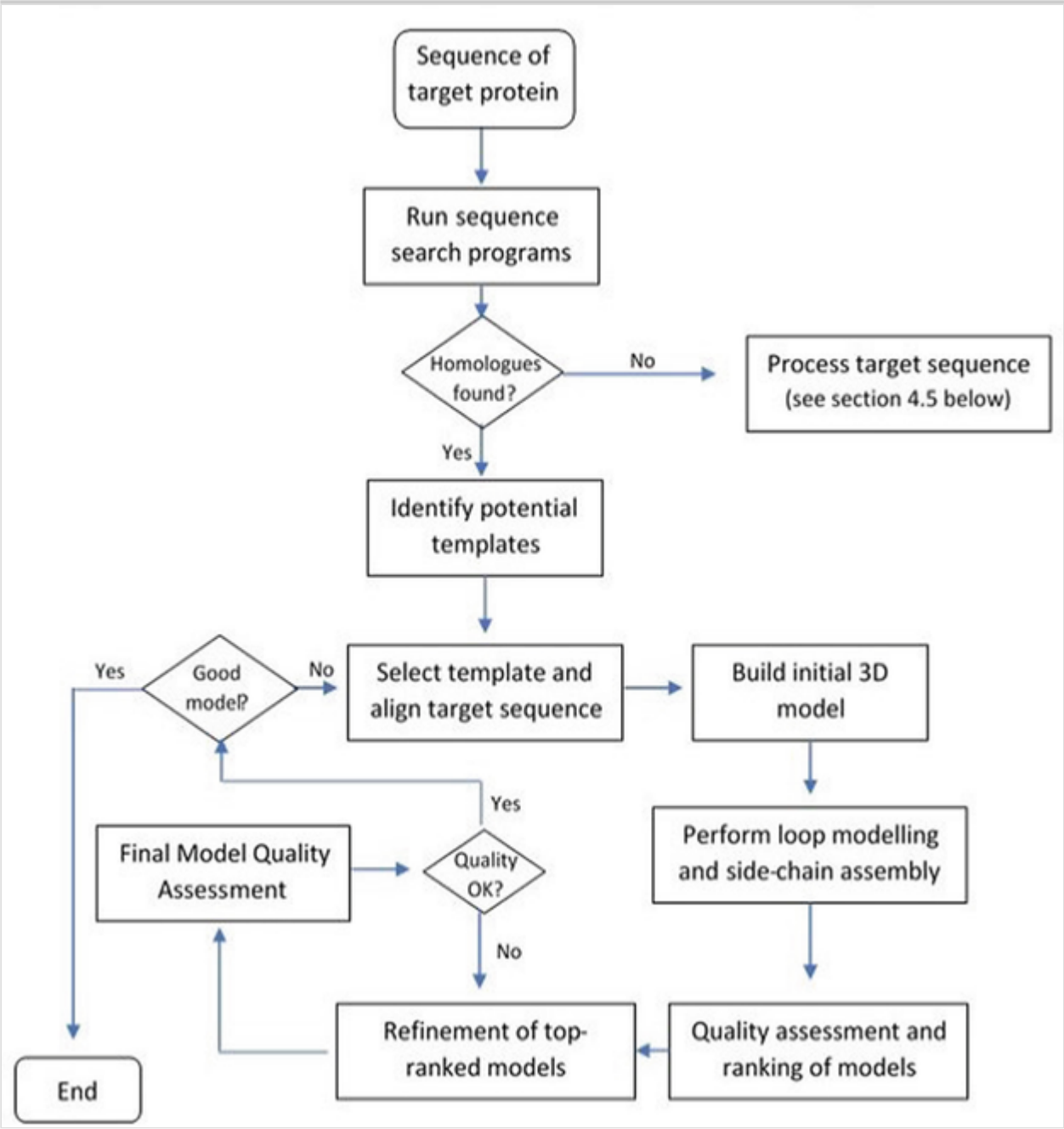
AQ6

## 4.4. Refinement

Refinement is the process of taking a raw model and attempting to improve its quality score by making small changes to the 3D structure in the hope and expectation that the newly produced model will be closer to the native protein than the original. Refinement programs essentially perform two separate functions; the first is one of sampling, that is, to create improved 3D models from those already built by the modeling software (often by MD employing the AMBER or CHARMM force fields) and the second is one of scoring these models, mostly via energy functions (such as DFIRE, RWPlus, and Rosetta), so that improvements can easily be identified [36]. It is in the second function that refinement programs overlap significantly with model quality assessment programs and the process of MQA and refinement can often be iterative as shown in Fig. 1 ~~above~~ below.

**Fig. 1**

A flow chart of the key stages in template-based modeling




As well as performing two functions, refinement programs can be broadly split into two types. First are those, sometimes referred to as manual programs, which perform very computationally intensive functions such as molecular dynamics (MD) and Monte Carlo statistical simulations and may also be augmented by applying knowledge-based constraints. These tend to be programs available to download and run locally in Linux or available to run from specialist research groups who complete in the CASP experiments. Second are the automated server-style programs that are available via public webpages. These tend to be quicker and focus more on computationally less-intensive methods such as side-chain optimization and less stringent energy minimization functions [40]. The second

group tend to make more conservative changes to the models, which is often desirable if the models are of reasonably good quality in the first place. Table 6 lists a number of publicly available refinement servers.

**Table 6**

Publicly available refinement webservers. (Reproduced from *Methods for the Refinement of Protein Structure 3D Models*, 2019 (Adiyaman R and McGuffin LJ) with permission from International Journal of Molecular Science [36])

| | | |
|---|---|---|
| PREFMD [41] | Developed by the Feig group, based on molecular dynamics (MD) | http://feiglab.org/prefmd |
| locPREFMD [42] | As above but focussed on local (per residue) quality | http://feig.bch.msu.edu/web/services/locprefmd/ |
| GalaxyRefine [43] | From the Seok group, focused on side-chain repacking | http://galaxy.seoklab.org/refine |
| KoBaMIN [44] | Energy minimization strategies using a knowledge-based force field | http://csb.stanford.edu/kobamin |
| Princeton TIGRESS 2.0 [45] | Combines many strategies from other servers, scored well in CASP experiments | http://atlas.engr.tamu.edu/refinement/ |
| ModRefiner [46] | Multi-step algorithm for side-chain optimization with physics and knowledge-based force fields | http://zhanglab.ccmb.med.umich.edu/ModRefiner |
| 3DRefine [47] | Optimization of H-bonds and energy minimization with physics and knowledge-based force fields | http://sysbio.rnet.missouri.edu/3Drefine/ |
| ReFOLD [48] | A quasi single-model approach with H-bond optimization and MD, using ModFOLD, from the IntFOLD server | http://www.reading.ac.uk/bioinf/ReFOLD/ |
| FG-MD [49] | MD-based algorithm using TM-align to identify analogous fragments from the PDB | http://zhanglab.ccmb.med.umich.edu/FG-MD/ |

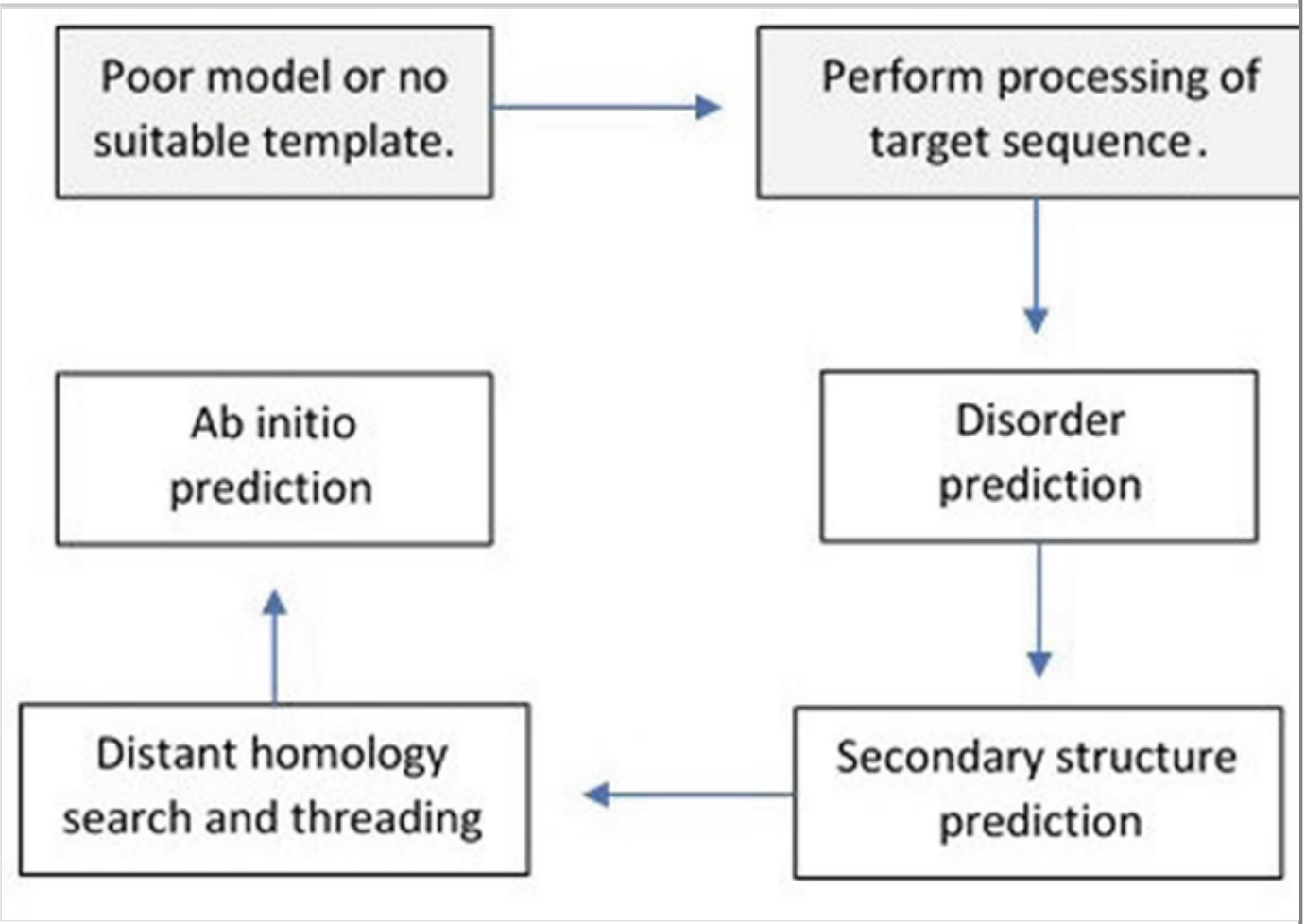
## 4.5. What if your Model Is Not a Good One?

If your model does not score well when subjected to quality assessment programs and attempted refinement, then it is likely that the template, on which it is based, is not a good match. Checking back to the flow chart in Fig. 1, we can see that problems may become obvious much earlier than this if there are few or no homologs identified for your target sequence. In either case, there are a number of avenues that may lead to an improvement in the model quality. These are

summarized in Fig. 2 below and the following sections where one or more options may be necessary.

**Fig. 2**

A flow chart showing some alternative modeling strategies




## 4.6. Disorder and Secondary Structure Prediction

One possible reason that your chosen modeling software fails to produce a good model of your target protein may be that it contains some intrinsically disordered regions (IDRs). Many proteins contain flexible regions in place of well-defined secondary structure [50], and these regions have been linked with a number of functions including recognition and binding of ligands and DNA, signaling and cell cycle control or even potential phosphorylation sites. In many of these cases, the phenomenon of disorder-to-order is only observed upon binding and so the protein, in its native-unbound state, will be unlikely to comply with programmed expectations of 3D structure. Disorder prediction may therefore give some clues as to why models are poorer than expected.

In a similar way, it may be worth checking the predicted secondary structure of your target. Although modern modeling software is very good at recognizing folds and domains that occur at different positions in polypeptide chains, there is the possibility that multi-domain proteins containing long loops and areas of disorder will be poorly scored and ranked with the available software. It is therefore worth checking secondary structure agreement between target protein and the templates and/or the models generated, to inform your interpretation of the models you are presented with. Indeed, McGuffin writing in 2010 [12] asserted that simple scores based on secondary structure compatibility can be very effective model quality assessment and be used to filter out models with incorrectly or poorly formed secondary structures. See Table 7 for a list of disorder prediction tools.

**Table 7**

Protein disorder and secondary structure prediction tools

| | | |
|---|---|---|
| JPred4 [11] | Secondary structure prediction online server | www.compbio.dundee.ac.uk/jpred/ |
| PSIPred [10] | Hosted by UCL, London. Secondary structure prediction with links to associated applications | http://bioinf.cs.ucl.ac.uk/psipred/ |
| Disopred [51] | Recognition of disordered regions | http://bioinf.cs.ucl.ac.uk/psipred/ |
| IUPred [50] | Predictions of intrinsically unstructured proteins | https://iupred2a.elte.hu/ |
| PrDOS [52] | Protein DisOrder prediction system | http://prdos.hgc.jp/cgi-bin/top.cgi |

## 4.7. Distant Homology Searches, Fold Recognition, and Threading Programs

In order to negate the limitations of sequence alignment, particularly where sequence identity is below that 35% threshold, the process of protein fold recognition was developed [1, 20]. This technique employs the rationale that evolutionary homologs often display less structural divergence than sequence divergence [35] and therefore less reliance on matching sequence and more on matching fold structure can result in less clutter of sequence-related but structurally distant template suggestions. Fold recognition commonly involves statistical methods (e.g., Hidden Markov Models—HMM) [20] to compare sequence profiles of targets with potential templates and identify the most suitable ones from which to construct 3D models. Traditionally, threading methods were also developed

which would fit or “thread” target sequences into the backbones of existing structures and then evaluate suitable templates using statistical energy potentials. Stand-alone individual fold recognition and threading techniques have enjoyed success in previous CASP experiments and include those listed below in Table 8. However, there is now a question as to whether their predictive powers have reached a plateau [57], as most successful servers now deploy a combination, or consensus, of alternative techniques.

**Table 8**

Tools when no close matches are found

| | | |
|---|---|---|
| THREADER [53] | Fold recognition methods for predicting protein structure | http://bioinf.cs.ucl.ac.uk/software_downloads/threader/ |
| GenTHREADER [54] | Rapid fold recognition, matching sequences against PDB chains assuming an evolutionary link | http://bioinf.cs.ucl.ac.uk/web_servers/ |
| pGenTHREADER [54] | Highly sensitive fold recognition using profile–profile comparison | http://bioinf.cs.ucl.ac.uk/web_servers/ |
| pDomTHREADER [54] | Highly sensitive homologous domain recognition using profile–profile comparison | http://bioinf.cs.ucl.ac.uk/web_servers/ |
| HHPred [34] | Tertiary structure prediction and threading, part of the HH-suite of programs | https://toolkit.tuebingen.mpg.de/tools/hhpred |
| MUSTER [55] | MUlti-sources ThreadER, a threading algorithm combining sequence profile–profile alignment with structural information | https://zhanglab.ccmb.med.umich.edu/MUSTER/ |
| SPARKS-X [56] | Fold recognition software | http://sparks-lab.org/yueyang/server/SPARKS-X/ |

## 4.8. Ab Initio or (Template) Free Modeling Methods

Ab initio modeling which is essentially synonymous with template-free modeling (TFM) is a technique that applies physics-based rules in order to estimate the structure of a target sequence using the sequence as the only input [18]. These programs do not query the PDB or any other database, instead relying on the application of physical algorithms to build the model from scratch.

The algorithms used will be very similar to those discussed so far, focusing on torsion angles, hydrophobicity, secondary structure agreement as well as energy

minimization and molecular dynamic technics. The computational power necessary to cope with the many degrees of freedom that present themselves in these cases is significant and many ab initio predictive servers run on either integrated CPU networks, powerful GPUs (graphical processing units), or neural networks and support vector machines (SVMs)—FALCON (a remote template alignment program employing a significant number of ab initio routines within its algorithms) harnesses the power of 20,000 volunteer CPUs for example [58]. QUARK represents a pure ab initio prediction methodology (there are others) whereas FALCON and Robetta (in the form of the upgraded ROSIE site—see Notes, Subheading 8) include a certain amount of ab initio routines behind the scenes while performing model building (see Table 9 for weblinks).

**Table 9**

A sample of available *Ab initio* or de novo modeling software

| | | |
|---|---|---|
| FALCON [58] | Software specializing in aligning query proteins with conserved regions | http://protein.ict.ac.cn/TreeThreader/ |
| QUARK [59] | Structure prediction and protein folding to construct 3D models from amino acid sequence only | https://zhanglab.ccmb.med.umich.edu/QUARK/ |
| ROSETTA [19, 28, 29] | ROBETTA server (robot-Rosetta) provides ab initio folding and structure prediction, as well as fragment selection | http://robetta.bakerlab.org/ |

QUARK is typical of many of the modern ab initio prediction sites which now tend to use small fragments (1–20 residues long) and reference their own fragment database [59].

Here, it might be prudent to briefly mention the recent development of TFM programs specializing in amino acid contact prediction. The two leading proponents of this technology are Google DeepMind, using the Alphafold algorithm, and DMPfold. Alphafold uses a system of contact distance and angle predictions that are then solved by gradient descent mathematics [60]. DMPfold works slightly differently by predicting inter-atomic distances, torsion angles, and main chain hydrogen bonding to drive the folding prediction. Both use powerful neural networks and have reported success with CASP tertiary structure targets; DMPfold predicted 56% of folds correctly in CASP13 targets and Alphafold led the field with 72% correct [61].

# 5. Comprehensive or Integrated Structure Prediction Webservers

The Swiss Institute of Bioinformatics (SIB) website (https://www.expasy.org/proteomics/protein_structure) has links to many publicly available programs designed to perform specific stages of the prediction process as well as those which perform the full service from start to finish. OmicX (https://omictools.com) is another useful website with an abundance of well-categorized resource links. It must also be mentioned that some of the above-mentioned server programs also offer complete sequence to 3-D model functionality or are part of a webserver suite or collection of programs designed to complement each other, for example, the UCL PSIPRED workbench (http://bioinf.cs.ucl.ac.uk/psipred/) allows one or many stages of the protein prediction pipeline to be undertaken at any one time with a simple tick-box system.

Below we will limit our focus to five leading one-stop webservers and describe briefly their mode of action and any advantages or special features they provide. They are listed in alphabetical order.

## 5.1. IntFOLD

IntFOLD is an integrated protein structure and function server consisting of a suite of interlinked programs developed by the McGuffin group and hosted by the University of Reading. As with many stand-alone servers, IntFOLD uses its own algorithms along with those from numerous other servers in order to multiply the power of template selection and accuracy of predicted models [21, 62].

INPUT: IntFOLD simply accepts the sequence of the target protein of interest. There is the option to provide a name for the job and an email address to which the results page link can be sent. Click on the *IntFOLD submission* link to be taken to the latest version of the program. If an email address is not submitted, users should be sure to bookmark or save the link to the results page as it will be lost upon navigation away from the page.

*MODE*: IntFOLD works broadly on a two-step process; first, is a single template modeling step with Accuracy Self Estimate (ASE) scoring followed by a second multiple template modeling step, again with ASE scoring.

The first step of template identification harnesses the power of 14 separate algorithms, six stand-alone fold recognition programs—SP3, SPARKS-2,

HHsearch, COMA, SPARKS-X, and CNFSearch, and the eight threading programs comprising the LOMETS package. Each individual algorithm may submit up to 10 templates (140 in total), which are then run through the IntFOLD server's clustering and scoring algorithm ModFOLDClust2.

The second step involves an iterative multi-template modeling (MTM) regime using the cluster scores to rank the templates found in **step 1**. Firstly, the top two alignments are used to construct an initial model, this is then compared to models made using the top ranked plus any other template, the best model is selected based on amino acid coverage of the models. This is performed twice more for the evolving model before selected models are re-scored with ModFOLDclust. The 4-stage iterative model building and comparison process is then repeated. Additionally, I-TASSER and HHPred [34] are used to build three models each and these are added to the model group from the iterative process which are then fed into a ranking and refinement loop. Using ModFOLD6_Rank [21] and reFOLD algorithms, models are continuously ranked and refined via molecular dynamics procedures and the final top five-ranked models from this cyclic process constitute the IntFOLD output.

*OUTPUT*: The output file lists the top five models ranked by global model quality score and accompanied by a color-coded p-value. The following sections are also included; Disorder prediction, Domain Boundary prediction, Binding site prediction, and full quality assessment results. These are comprehensively described and explained on the IntFOLD Webserver help page (https://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD_help.html#examples) and so will not be repeated here. Users may download the data files for the predictions via the hyperlinks on the results page.

*CREDENTIALS*: In CAMEO server benchmarking IntFOLD4 was rated second on the common subset comparison (1-year performance 2016–17) and IntFOLD5 was rated first in 3-D data results for 3 months (Oct 2018–Jan 2019). The McGuffin group has also been competitively ranked in numerous recent CASP experiments [4].

## 5.2. I-TASSER

Developed and administered by Zhang Lab of the University of Michigan, the acronym stands for Iterative Threading ASSEmbly Refinement [22].

*INPUT*: In addition to the basic sequence in FASTA format, I-TASSER allows users to specify additional restraint data if known, for example, distance restraints in the form of atom contacts. If users would like to specify particular proteins to be used as homologs, their PDB codes can be entered and there is also the facility to upload a complete 3D homolog structure in PDB file format should that be required. Users can also take advantage of TASSER's threading credentials by excluding close sequence homologs and going below the usual cut-off of 25% sequence identity.

MODE: I-TASSER is a suite of programs. The initial fold recognition is carried out by the LOMETS meta-server with subsequent fragment threading by the MUSTER [55] algorithm. The fragments are then assembled into potential models with loop sections built by ab initio methods as necessary. SPICKER then selects the best models by clustering on a lowest energy basis and the process is verified by parallel model-build using TM-Align. The models are then re-clustered, and the final model is constructed using REMO software.

*OUTPUT*: Submissions can take 1–2 days to run by the end of which users will be emailed a results webpage link. The results are extensive and include a secondary structure visual display, solvent accessibility display, and a B-factor graph showing variation along the mode (*see* **Note 1**). Following this is an interactive list of the templates used as well as the top five models viewable in a JMol-style graphical user interface. Each of the model files is downloadable and accompanied by a C-score, TM score, and RMSD. Included at the bottom of the results page are some potentially useful sections on predicted co-factors and binding sites, enzyme potential data, and gene ontology information.

*CREDENTIALS*: I-TASSER was ranked as the top server in CASP 7, 8, 9, and 10.

## 5.3. Phyre2

This is an updated version of the Phyre server that has been completely rewritten with the emphasis on both enhanced technical attributes and usability. The acronym stands for Protein Homology/analogY Recognition Engine V 2.0 and is run by the Structural Bioinformatics Group at Imperial College, London, making up part of the Genome3D collaboration between UCL, Imperial, Cambridge, and Bristol universities [20].

INPUT: Phyre2 can be accessed from the Phyre2 homepage, which will accept a sequence in FASTA format as well as an email address for results. It can also be accessed via the Genome3D page (http://genome3d.eu) where a FASTA, keyword

or UniProt id submission returns a list of matches that, upon selection, lead to a predicted domains page. Here there are links to CATH and SCOP for protein classification information and Phyre2 for 3-D modeling (as well as links to some other Genome3D annotation software).

*MODE*: As with many servers, Phyre2 makes use of a number of other programs. Alignment and template detection is now upgraded from a PSI-BLAST search to a HMM-based fold library scan using HHsearch/HHpred software. Secondary structure is also predicted using PSIPRED. Phyre2 has a sophisticated mechanism for the management of insertions, deletions, and disordered or missing loop regions; employing a fragment-matching library and testing dihedral angle and energy scores to ensure the lowest possible perturbance in the structure as potential fragments are inserted. There is also an acknowledgment of the persistent problem of few templates or templates that only match one domain for a multi-domain target. Here the *ab initio* modeling software Piong, which is designed to work as a virtual ribosome, is employed to build as much of the model as necessary. Lastly, DISOPred software predicts areas of disorder and the R3 protocol uses a rotamer library to orientate amino acid side chains.

*OUTPUT*: Results are emailed to users with a link to the results page. The page is split into four sections; firstly, a model based on the top-ranked template which can be viewed interactively in JMol; secondly, a detailed graphic of predicted secondary structure and potential disorder scores; third is nice graphic of all templates and the percentage alignment for each, these are interactive and link to the fourth section below which lists all templates' structures and PDB information. These are downloadable individually, and there is a Download as zip option for the whole results page (*see* **Note 2**).

*CREDENTIALS*: Phyre2 is an older server that was been ranked sixth in CASP9 and tenth in CASP10. However, the authors are keen to point out that there is only 2–3% difference other servers' performance (measured by GDT_TS) [20], (with the exception of I-TASSER which scored slightly better in cases where only remote homologs exist).

## 5.4. Robetta

Robetta is the public-facing webpage of the Rosetta server prediction program developed by the Baker lab at the University of Washington, USA, and now administered by the Rosetta Commons group. Rosetta has a long history as a

competitor in CASP and Robetta is a free-to-use front end-running the powerful Rosetta algorithms that have been so successful [19, 28, 59].

INPUT: Users must register in order to run jobs on Robetta. There are essentially three options upon registration; Rosetta comparative modeling (CM), Rosetta *ab initio* modeling (AB), or a fully automated pipeline. Users can paste (FASTA) or upload an amino acid sequence and also upload templates or alignments of their own if required. It is also possible to add custom distance constraints, if known. Users are only allowed one job at a time and jobs are run on a two-stage process; firstly, the identification of templates and secondly domain 3-D modeling. Users will be required to pick a domain to model after stage one and may submit only one domain at a time to conserve computing power (*see* **Note 3**).

*MODE*: Robetta essentially runs four separate algorithms for template selection and alignment; these are RaptorX, HHPred, SPARKS-X, and Map align. As above, users are able to upload their own templates and alignment data if they which to bypass this stage. Rosetta algorithms then perform 3-D modeling on a domain by domain basis and also check potential interface areas by Alanine scanning (each amino acid is in-turn replaced by Alanine and the effect on the calculated binding energy computed) for binding and interaction prediction.

*OUTPUT*: Jobs typically take 1–2 days to run and users receive access to the results page via email. The results are comprehensive and include a multi-server secondary structure annotation with disorder predictions plus interactive RasMol annotations of the top five models, which can be colored by error estimation. Graphical error plots of distances (in Å) between Cα atoms of the model compared to the native structure also accompany each model. The results page is interactive and a click on each domain will reveal the templates and alignments used to build it as well as a cluster graph showing its position relative to the average. For comparative modeling, a predicted confidence value equivalent to GDT_TS is provided. For ab initio modeling, a predicted confidence value equivalent to TM-score of the top 10 Rosetta scoring models is provided instead.

*CREDENTIALS*: Robetta has competed in CAMEO since 2014 and cites its success in terms of LDDT score (Local Distance Difference Test—which evaluates inter-atomic distances). Robetta averages around 69 (0–100 where higher scores are better). The error estimates included in results are also evaluated through CAMEO and Robetta achieves an average model confidence score of 0.85.

### 5.5. Swiss-Model

This was the first fully automated server developed over 20 years ago and is now a comprehensive website with enhanced functionality administered by the Swiss Institute of Bioinformatics (SIB) [16].

*INPUT*: As well as a FASTA sequence users can input the UniProt accession code for the target. There also exists the facility to upload potential template files, but familiarity with the SIB Swiss-PDBViewer, also known as DeepView, will likely be necessary for this.

*MODE*: There are a number of key features to SWISS-MODEL. It is designed specifically to run HMM modeling, via HHblis [34] software, on the SWISS-MODEL Template Library (STML); an amalgamated version of the SWISS PROT and PDB databases augmented with derived data allowing the differentiation between bound ligands and solvent molecules. SWISS-MODEL will also run a BLAST search and check secondary structure via PSIPRED before allowing the user a choice between automated or manual selection of the templates found. If manual mode is selected, the templates are listed along with their Global Mean Quality Estimation score (GMQE—essentially an average of QMEAN [39] scores applied to each individual amino acid) and information on predicted ligands, oligomeric state, and sequence alignment. Users are able to select any number of templates and these are then displayed in a 3-D structural super-position as well as a 2-D cluster graph of evolutionary distance. Users can then choose their potential templates based on clustering, domain matches, and sequence identity scores.

SWISS-MODEL will then build an all-atom model using ProMod II software with a back-up comparison built using MODELLER [23].

OUTPUT: Users get a comprehensive listing of model coordinates, target-template alignment, step-by-step modeling log, information on potential oligomeric state, potential ligands, and co-factors as well as a QMEAN score, all of which can be downloaded. The models within the graphical interface are also colored by QMEAN to show areas of higher and lower confidence.

## 6. CASP and CAMEO

To give some context to the programs and rating credentials presented in Subheading 5, it is worth expanding here on the CASP and CAMEO community-

wide experiments (first referenced in the introduction) which form the arena in which modeling expertise is tested and advanced.

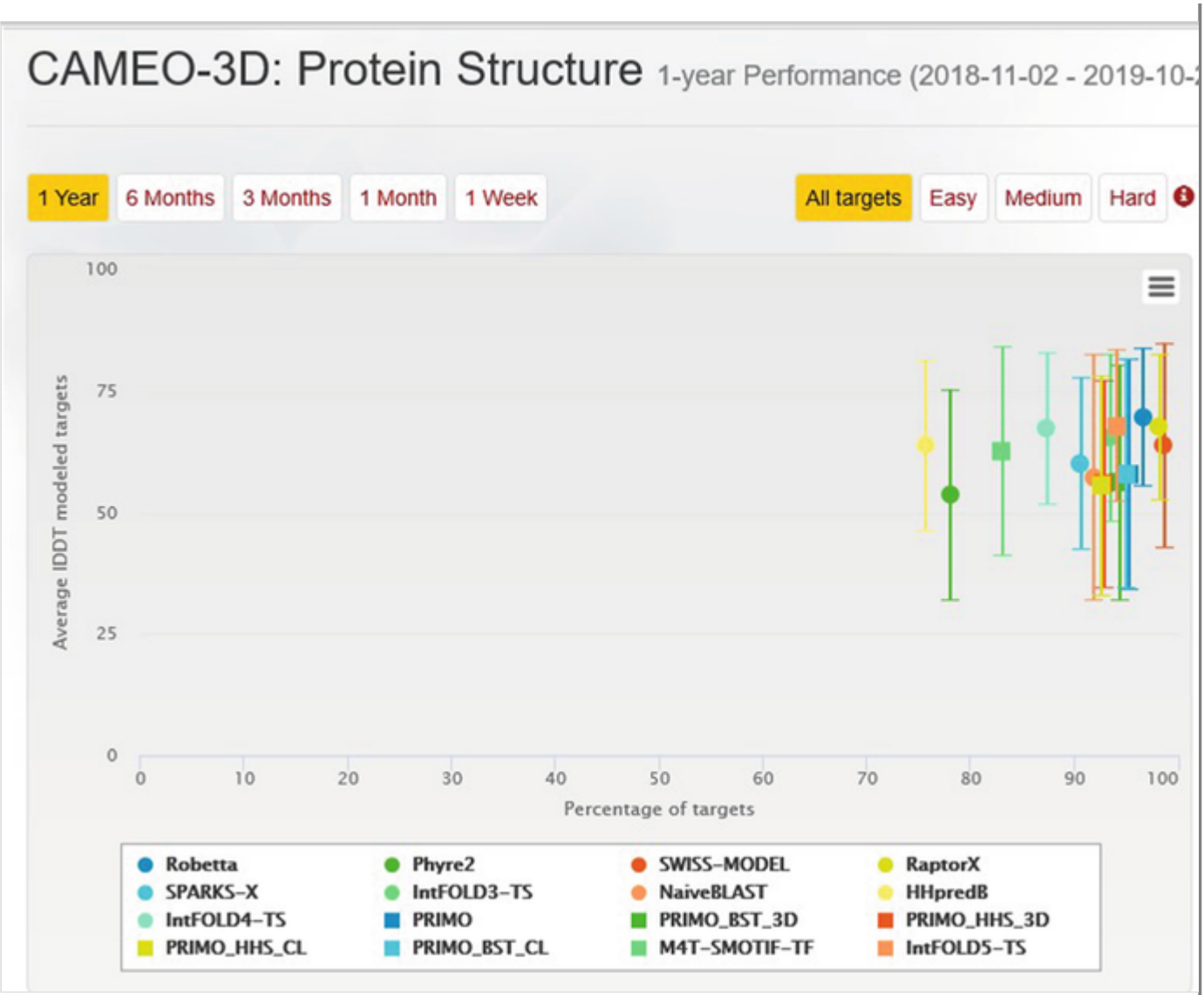
The CASP experiment has been running as a biannual blind tertiary structure prediction competition since its inception by John Moult and associates in 1994 [63]. The purpose has been to provide a vehicle for the objective assessment of the prediction capability of in silico groups globally with the added benefit of shared practice and identification of technical advancement. Organizers source soon-to-be-solved crystal or NMR 3-D structures from researchers and invite in silico prediction groups to solve the structure before revealing the answers and scoring groups' efforts around 9 months later [35]. These experiments have seen the discipline of in silico protein structure prediction rise in integrity over the past 25 years with CASP1 attracting 35 invited predictor groups [63] compared to CASP6, (run 10 years later in 2004) which received over 30,000 predictions from 200 predictor teams [35] and CASP8 (2008) representing peak predictor participation with 253 groups across 24 countries worldwide [64].

Since the time of its inception to the latest version the focus of the CASP experiment has changed and expanded from mostly ab initio modeling to comparative methods (TBM) which are able to exploit the wealth of structural information now available (by CASP10 (2012) there were 1393 distinct folds available in the PDB and a total of 87,000 solved protein structures [65].

CAMEO (Continuous Automated Model EvaluatiOn—see Fig. 3) is a server-based experiment run along similar lines to CASP but differing in that participating servers must be fully automated with no human intervention in the prediction process. Servers receive their targets on a weekly basis and have 3 days in which to complete the prediction and return results to CAMEO. The ratings and metrics on the relative successes of the servers is a good indication of their competitiveness and likelihood of providing a good quality model.

**Fig. 3**

A screenshot from the CAMEO website showing participating servers. (Taken from ht tps://www.cameo3d.org/sp/1-year/difficulty/all/?to_date=2019-10-26)




# 7. Protein–Protein Interactions (PPI) and Quaternary Structure Prediction

While both CASP and CAMEO experiments include predictions of the interaction of proteins to form dimers and some higher level oligomers, the third community-wide prediction competition CAPRI (Critical Assessment of Prediction of Interactions) forms the area of expertise in PPI and quaternary structure prediction. However, communities are now merging somewhat with CASP 11 (2014) and CASP 12 (2016) seeing joint CASP-CAPRI collaborations on many prediction targets, representing a crossover of docking and homology modeling expertise.

## 7.1. Docking Programs

Many program routines currently used in the CAPRI experiment were originally developed to predict protein docking interfaces with either ligands or with

themselves to form homodimers which explains the inclusion of the word “dock” in many program names. Although these programs can often perform a protein–ligand docking function, the ones listed here have been developed to focus primarily on protein–protein interactions. If a program specifically for docking is required, a popular choice is Autodock Vina.

A number of different docking approaches have been developed to predict protein–protein interactions. A favorite technique is the use of a Fast Fourier Transform (FFT) to search all possible binding modes in a 6-D search space (3 rotational and 3 translational) [66] but there are others based on shape complementarity, spherical harmonics, and identification of Zernike shape descriptors as well as those employing more traditional physics-based measurements such as energy minimization, side-chain orientation, and solvent accessibility.

See Table 10 for a list and brief description of some of the main players in the prediction of interactions and quaternary structure via docking algorithms.

**Table 10**

Docking-based PPI modeling software

| | |
|---|---|
| GRAMM-X [66], ZDOCK [67], and MEGADOCK [68] | Fast Fourier transform (FFT)-based programs |
| FRODOCK [69] | Modified FFT technique (Chacon et al., 2009), using a reduced 3-D search space to save time and computer power yet reportedly achieving a comparable level of accuracy |
| PatchDock [70] | Uses image segmentation techniques to map the contours of the surface of a protein followed by shape complementarity and symmetry to fit the protein surfaces together |
| Hex [71] | Uses spherical harmonics (D. Ritchie) |
| RosettaDock [28] | Uses a combination of side-chain orientations and free-energy calculations linked to its probability-based Monte Carlo algorithm |
| LZerD [72] | A unique approach identifying Zernike 3-D shape descriptors followed by complementarity calculations |
| ClusPro [73] | Models are clustered together depending on the location of the interface residues, the logic being that the size of clusters is proportional their probability of representing the native model |
| HADDOCK [74] | A physics-based scoring function based on a combination of van der Waal’s interactions, electrostatics, and desolvation measures |

All approaches have had success over the rounds of CAPRI experiments with ClusPro scoring a success rate of 5 high and 3 medium quality models, followed by HADDOCK with 4 high and 1 medium (from 12 targets) in 2009 and LZerD scoring 4 high and 3 medium models from 20 targets in 2016 (data from the server modeling section of CAPRI [73]). RosettaDock has also enjoyed success, predicting all 5 small targets with medium to high accuracy in rounds 3–5 [28] as well as being ranked second in the 2014 predictor server rankings [75]. All servers are listed with their varying levels of success in the 2014 CAPRI round 30 [75] at htt p://www.capri-docking.org/resources/#performance-of-docking-servers-in-capri. It must be added that most success in protein interaction prediction has come in the form of predicting dimers and certain higher order oligomers exhibiting spherical symmetry with hetero complexes continuing to present problems [76]. Analysis of the joint CASP/CAPRI experiments by Lensink et al. (2016) [75] suggests that, in general, docking approaches to predicting quaternary structures performed better than template-based modeling due, in part, to the increased difficulty of finding reliable oligomeric crystal templates in the PIR database. Therefore, although an increasing number of 3-D modeling programs will offer a likely quaternary structure for a target sequence it may be worth bearing in mind the additional difficulties that this process involves when considering the accuracy of the final model.

## 7.2. The Evolution of Docking Methods

Although docking programs can produce very good models of homodimers, they are less well adapted to identifying quaternary structure straight from sequence especially for hetero or larger complexes. While some of the programs listed above have been adapted to predict higher level homomers, e.g., MZDock and MultiLZerD (as demonstrated by Nakamura et al. (2017)) [77], their use often still requires a catalog of specialist software and results can be variable. One server to both beef-up its computing power and allow easy user input directly from a webpage interface is MEGADOCK 4.0 (accessible as MEGADOCK-Web http://ww w.bi.cs.titech.ac.jp/megadock-web/).

Other specialist quaternary prediction sites that are publicly available via a webpage and require only sequence data in FASTA format as input include SWISS-MODEL, QuaBingo, and Galaxy.

Bertoni et al. (2017) [78] reported their attempt to go from sequence straight to quaternary structure using SWISS-MODEL that samples multiple template

databases as well as adding a co-evolution distance measure score—termed PPI fingerprint. If it is considered possible to build a quaternary model using SWISS-MODEL, the quaternary structure quality estimate (QSQE) score will be included in the output.

Another study, Tung et al. (2016) [79] reported their description of the program QuaBingo that identifies conserved domains using the BLOCKS database of motifs based on SWISSPROT. QuaBingo also adds a pseudo amino acid descriptor (PseACC) that takes into account the hydrophobic-hydrophilic character of individual residues. QuaBingo can be accessed from http://predictor.nchu.edu.tw/QuaBingo.

Galaxy also has a homomer prediction facility based on a simple FASTA sequence submission (http://galaxy.seoklab.org/) as Galaxy-Homomer.

# 8. Notes

1. When using I-TASSER:

   Models are selected by clustering and although there is good evidence that clustering improves model identification [37], care should be taken when a target sequence has few homologs as clustering may be less powerful. Also, the ranking of the models by cluster size presents the potential for a good model (higher C-score) being omitted from the top of the models list as it appears in a smaller cluster. Results should be checked for these issues.

2. When using Phyre 2:

   Phyre2 has a number of ad-on functions that may be useful.

   BackPhyre is a genome search tool allowing users to search for homologs to their solved structure in specific genomes.

   One to one threading can be used if users have biological information indicating that a specific protein should be used as the template. A file can be uploaded.

   Phyre Alarm is a scanning service which checks fold libraries on a weekly basis and updates users who have not found a good template match in their initial modeling attempt.

Phyre Investigator give access to extra information on model quality analysis, alignment confidence, and Ramachandran analysis as well as catalytic site, mutation analysis, and potential interface detection.

Lastly, users can opt for Batch Analysis, where up to 100 jobs can be scheduled to run automatically and Job Manager that gives access to a page with all previously run jobs.

3. When using Robetta:

Rosetta software is available to download if users would prefer to run the algorithm locally from the command line. There is also an option to download pyRosetta for those interested in running the software via Python. From the Robetta homepage are links to the latest Rosetta incarnation called ROSIE. This has links to a whole host of functional characterization programs (one could say a whole lotta Rosie!) and would be worth visiting.

A list of scoring functions often encountered in protein structure prediction is given in Table 11.

**Table 11**

A list of scoring functions often encountered in protein structure prediction

| **Predictive scores (for model quality assessment)** | |
|---|---|
| C-score | (I-TASSER). This is a confidence score calculated for threading template alignments. Scores range from −5 to 2 with higher scores indicating a better alignment |
| E-value | (BLAST and RAPTOR). Related to p-value, for two sequences with n alignments, E-value represents the expected number of false alignments having greater than n correctly aligned positions. The closer to 0 the better |
| LG score | (PCons). Essentially a *p*-value for the significance of a structural similarity match. A significant threshold would be $1 \times 10^{-1.5}$ (0.031), so anything below this figure would represent a potentially good match between a model and the target |
| MaxSub score | Identifies the largest set of Cα atoms that superimpose well over two structures so focusing on well-predicted regions. Produces a score between 0 and 1 with 1 being the best, normalised for the size of the overlap so that larger sequences do not automatically score better than shorter ones |
| ProQ score | (PCons). This is the -log of LG score, e.g., for a significant LG score of $1 \times 10^{-1.5}$ The ProQ score would be 1.5. Therefore, 1.5 and upwards are good scores |
| *p*-Value | The proportion of models with a particular score that do not share any |

| | |
|---|---|
| | similarity with the native structure, i.e., will have the same alignment purely by chance. <0.001 = 1/1000 chance (or less) that the model is incorrect; <0.01 less than a 1/100 chance; <0.05, less than a 1/20; <0.1 less than a 1/10; >0.1 likely to be a poor model with little or no similarity to the native structure |
| Qmean score (qualitative model energy analysis) | The simplest form of this, Qmean4, is the sum of four measures; geometric analysis of the torsion angles of the carbon backbone, CB interactions, all atom interactions, and a solvation score (QMean6 additionally includes a secondary structure agreement score and a solvent accessibility agreement as percentages. A Qmean4 of 1 is good with 0 considered acceptable but, as with Z-score, a negative figure indicates a poorer fit. Qmean scores are often transformed into Z-scores for ease of comparison with experimentally determined structures |
| S-score | (PCons). A global super-position score calculated as a transformation of RMSD on a per amino acid residue basis. 1 would represent a perfect score and 0 a useless model |
| TM-score | This is a measure of the similarity of two protein structures based on a weighted RMSD score, i.e., small RMSD values are weighted more strongly than large scores in an attempt to overcome the distortion of RMSD for good models with local errors. Scores can range from 0 to 1 with >0.5 representing a strong match and < 0.17 a match no better than random |
| Z-score | A Z-score is an expression of the number of standard deviations from the mean structure of the templates. A Z-score of zero would indicate that a template represents the mean structure, a negative score would indicate a worse fit than the mean whereas a positive score would indicate a better fit. However, it must be remembered when dealing with normal distributions and standard deviations, the further one travels from the mean, in any direction, the more likely one is to be looking at an outlier and the true value is likely to be close to the mean |
| **Observed scores (obtained when a model is compared to the true structure)** | |
| Global model quality score | The global model quality scores range between 0 and 1. In general, scores less than 0.2 indicate there may be incorrectly modeled domains and scores greater than 0.4 generally indicate more complete and confident models, which are highly similar to the native structure |
| GDT_TS (Global distance test total score) | A CASP observed score. Explanations may be found at http://predictioncenter.org/casp13/doc/help.html#GDT_TS |
| B-factor | Often known as a temperature factor, this measurement is traditionally supplied with crystallographic structures as a measure of the displacement of individual atoms from their true position. Measured in angstroms squared, 0 would be a perfect score with anything below 30 $Å^2$ considered as acceptable and anything greater than 60 $Å^2$, questionable (for reference a 15 $Å^2$ score would equate to a mean displacement of an atom by 0.44 Å and 60 $Å^2$, a mean displacement of 0.87 Å) |
| RMSD (root mean square deviation) | This usually refers to the average distance of all amino acid pairs in two compared structures. Some programs will give a global score for the whole structure whereas others my give local scores per amino acid residue. Measured in Å, a good score would be <2.0 although this will depend on |

| | |
|---|---|
| | the resolution of the templates used to calculate the model. This measure, although widely quoted, is particularly sensitive to the problem of local alignment error discussed below |

# References


1. Rangwala H, Karypis G (2010) Introduction to protein structure prediction. In: Rangwala, Karypis (eds) Introduction to protein structure prediction: methods and algorithms. John Wiley & Sons

2. Cao R, Bhattacharya D, Adhikari B, Li J, Cheng J (2015) Large-scale model quality assessment for improving protein tertiary structure prediction. Bioinformatics 31(12):i116–i123

3. Berman HM, Westbrook J, Feng Z et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242

4. McGuffin LJ, Adiyaman R, Maghrabi A et al (2019) IntFOLD: an integrated web resource for high performance protein structure and function prediction. Nucleic Acids Res 47:W408–W413

5. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515

6. El-Gebali S et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432

7. Andreeva A, Howorth D, Chothia C et al (2014) SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res 42:D310–D314

8. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I (2017) CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res 45(D1):D289–D295

9. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

10. Jones DT (1999) Protein secondary structure prediction based on position-

specific scoring matrices. J Mol Biol 292:195–202

11. Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res 43:W389–W394

12. McGuffin LJ (2010) Model quality prediction. In: Rangwala, Karypis (eds) Introduction to protein structure prediction: methods and algorithms. John Wiley & Sons
AQ7

13. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774–797

14. Zahn-Zabal M, Michel PA, Gateau A, Nikitin F et al (2020) The neXtProt knowledgebase in 2020: data, tools and usability improvements. Nucleic. Acids Res 48(D1):D328–D334

15. Szklarczyk D, Gable A, Lyon D et al (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47:D607–D613

16. Biasini M, Bienert S, Waterhouse A et al (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252–W258

17. Buenavista M, Roche D, McGuffin LJ (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. Bioinformatics 28:1851–1857

18. Guo J, Ellrott K, Xu Y (2008) A historical perspective of template-based protein structure prediction. In: Zaki, Bystroff (eds) Protein structure prediction, 2nd edition, methods in molecular biology, vol 413. Springer

19. de Oliveira HP, Shi J, Deane C et al (2015) Building a better fragment library for de novo protein structure prediction. PLoS One 10:e0123998

20. Kelley LA, Mezulis S, Yates CM et al (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858

21. McGuffin LJ, Shuid AN, Kempster R et al (2018) Accurate template-based modeling in CASP12 using the IntFOLD4-TS, ModFOLD6, and ReFOLD methods. Proteins 86:335–344

22. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER suite: protein structure and function prediction. Nat Methods 12:7–8

23. Webb B, Sali A (2016) Comparative protein structure modeling using modeller. Curr Protoc Bioinformatics 54, John Wiley & Sons, Inc.:5.6.1–5.6.37

24. Wang Z, Eickholt J, Cheng J (2010) MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 26:882–888

25. Wallner B, Elofsson A (2005) Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics 21:4248–4254

26. Yachdav G, Kloppmann E, Kajan L et al (2014) PredictProtein—an open resource for online prediction of protein structural and functional features. Nucleic Acids Res 42:W337–W343

27. Ma J, Wang S, Zhao F, Xu J (2013) Protein threading using context-specific alignment potential. Bioinformatics (Proceedings of ISMB 2013) 29(13):i257–i265

28. Park H, Kim D, Ovchinnikov S, Baker D (2018) Automatic structure prediction of oligomeric assemblies using Robetta in CASP 12. Proteins 86:283–291

29. Simons K, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225

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

31. Larkin MA, Blackshields G, Brown NP et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

32. Sievers F, Wilm A, Dineen DG et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. Mol Syst Biol 7:539

33. Madeira F, Park YM, Lee J et al (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636–W641

34. Zimmermann L, Stephens A, Nam SZ et al (2018) A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J Mol Biol 430(15):2237–2243

35. Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15:285–289

36. Adiyaman R, McGuffin LJ (2019) Methods for the refinement of protein structure 3D models. Int J Mol Sci 20:2301

37. McGuffin LJ, Roche DB (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics 26:182–188

38. Uziela K, Shu N, Wallner B, Elofsson A (2016) ProQ3: improved model quality assessments using Rosetta energy terms. Sci Rep 6:33509

39. Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27:343–350

40. Feig M (2017) Computational protein structure refinement: almost there, yet still so far to go. Wiley Interdiscip Rev Comput Mol Sci 7:e1307

41. Heo L, Feig M (2018) PREFMD: a web server for protein structure refinement via molecular dynamics simulations. Bioinformatics 34:1063–1065

42. Feig M (2016) Local protein structure refinement via molecular dynamics simulations with locPREFMD. J Chem Inf Model 56:1304–1312

43. Heo L, Park H, Seok C (2013) GalaxyRefine: protein structure refinement driven by side-chain repacking. Nucleic Acids Res 41:384–388

44. Rodrigues JPGLM, Levitt M, Chopra G (2012) KoBaMIN: a knowledge-based minimization web server for protein structure refinement. Nucleic Acids Res 40:323–328

45. Khoury GA, Smadbeck J, Kieslich CA et al (2017) Princeton_TIGRESS 2.0: high refinement consistency and net gains through support vector machines and molecular dynamics in double-blind predictions during the CASP11 experiment. Proteins Struct Funct Bioinform 85:1078–1098

46. Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 101:2525–2534

47. Bhattacharya D, Cheng J (2013) 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. Proteins 81:119–131

48. Shuid AN, Kempster R, McGuffin LJ (2017) ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates. Nucleic Acids Res 45:W422–W428

49. Zhang J, Liang Y, Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 19:1784–1795

50. Dosztányi Z (2018) Prediction of protein disorder based on IUPred. Protein Sci 27:331–340

51. Jones DT, Cozzetto D (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31:857–863

52. Ishida T, Kinoshita K (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res 35:W460–W464

53. Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358:86–89

54. Lobley A, Sadowski MI, Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and

superfamily discrimination. Bioinformatics 25:1761–1767

55. Wu S, Zhang Y (2008) MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72:547–556

56. Yang Y, Faraggi E, Zhao H, Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. Bioinformatics 27:2076–2082

57. Skolnick J, Zhou H (2017) Why is there a glass ceiling for threading based protein structure prediction methods? J Phys Chem B 121:3546–3554

58. Wang C, Zhang H, Zheng W-M et al (2015) FALCON@home: a high-throughput protein structure prediction server based on remote homologue recognition. Bioinformatics 32:462–464

59. Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80:1715–1735
AQ8

60. Hutson M (2019). AI protein-folding algorithms solve structures faster than ever. Deep learning makes its mark on protein-structure prediction. In: Nature NEWS, ISSN 1476–4687. https://www.nature.com/articles/d41586-019-01357-6. Accessed 31 Oct 2019

61. Greener J, Kandathil S, Jones D (2019) Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat Commun 10:3977

62. Nealon J, Philomina L, McGuffin L (2017) Predictive and experimental approaches for elucidating protein–protein interactions and quaternary structures. Int J Mol Sci 18:2623

63. Moult J et al (1995) A large-scale experiment to assess protein structure prediction methods. Proteins 23:ii–iv

64. Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins 79(Suppl 10):1–5

65. Moult J, Fidelis K, Kryshtafovych A et al (2014) Critical assessment of methods of protein structure prediction (CASP) — round x. Proteins 82:1–6

66. Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34:W310–W314

67. Pierce BG, Wiehe K, Hwang H et al (2014) ZDOCK server: interactive docking prediction of protein-ProteinComplexes and symmetric multimers. Bioinformatics 30:1771–1773

68. Hayashi T, Matsuzaki Y, Yanagisawa K et al (2018) MEGADOCK-Web: an integrated database of high-throughput structure-based protein-protein interaction predictions. BMC Bioinformatics 19:62

69. Garzon JI, Lopéz-Blanco JR, Pons C et al (2009) FRODOCK: a new approach for fast rotational protein-protein docking. Bioinformatics 25:2544–2551

70. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33:W363–W367

71. Macindoe G, Mavridis L, Venkatraman V et al (2010) HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res 38:W445–W449

72. Peterson LX, Kim H, Esquivel-Rodriguez J et al (2017) Human and server docking prediction for CAPRI round 30-35 using LZerD with combined scoring functions. Proteins 85:513–527

73. Vajda S, Yueh C, Beglov D et al (2017) New additions to the ClusPro server motivated by CAPRI. Proteins 85:435–444

74. Vangone A, Rodrigues JP, Xue LC et al (2017) Sense and simplicity in HADDOCK scoring: lessons from CASP-CAPRI round 1. Proteins 85:417–423

75. Lensink M, Velankar S, Kryshtafovych A et al (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. Proteins 84(Suppl 1):323–348

76. Lafita A, Bliven S, Kryshtafovych A et al (2018) Assessment of protein assembly prediction in CASP12. Proteins 86:1–399

77. Nakamura T, Oda T, Fukasawa Y, Tomii K (2018) Template-based quaternary structure prediction of proteins using enhanced profile-profile alignments. Proteins 86(Suppl 1):274–282

78. Bertoni M, Kiefer F, Biasini M et al (2017) Modelling protein quaternary structure of homo- and heterooligomers beyond binary interactions by homology. Sci Rep 7:10480

79. Tung C-H, Chen C-W, Guo R-C et al (2016) QuaBingo: a prediction system for protein quaternary structure attributes using block composition. Biomed Res Int 2016:9480276