SWISS-MODEL: homology modelling of protein structures and complexes

Andrew Waterhouse¹,²,†, Martino Bertoni¹,²,†, Stefan Bienert¹,²,†, Gabriel Studer¹,²,†, Gerardo Tauriello¹,²,†, Rafal Gumienny¹,², Florian T. Heer¹,², Tjaart A. P. de Beer¹,², Christine Rempfer¹,², Lorenza Bordoli¹,², Rosalba Lepore¹,² and Torsten Schwede¹,²,*

¹Biozentrum, University of Basel, Klingelbergstrasse 50–70, CH-4056 Basel, Switzerland and ²SIB Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Klingelbergstrasse 50–70, CH-4056 Basel, Switzerland

ABSTRACT

Homology modelling has matured into an important technique in structural biology, significantly contributing to narrowing the gap between known protein sequences and experimentally determined structures. Fully automated workflows and servers simplify and streamline the homology modelling process, also allowing users without a specific computational expertise to generate reliable protein models and have easy access to modelling results, their visualization and interpretation. Here, we present an update to the SWISS-MODEL server, which pioneered the field of automated modelling 25 years ago and been continuously further developed. Recently, its functionality has been extended to the modelling of homo- and heteromeric complexes. Starting from the amino acid sequences of the interacting proteins, both the stoichiometry and the overall structure of the complex are inferred by homology modelling. Other major improvements include the implementation of a new modelling engine, ProMod3 and the introduction a new local model quality estimation method, QMEANDisCo. SWISS-MODEL is freely available at https://swissmodel.expasy.org.

INTRODUCTION

Three-dimensional structures of proteins provide valuable insights into their function on a molecular level and inform a broad spectrum of applications in life science research. Often, complexes of proteins are central to many cellular processes. A detailed description of their interactions and the overall quaternary structure is essential for a comprehensive understanding of biological systems, how protein complexes and networks operate and how we can modulate them (1,2). Given their biological relevance, it is not surprising that the number of large complexes deposited per year in the Protein Data Bank (PDB) is growing rapidly (3). A significant contribution to this trend originates from the continuous progress of structure determination technologies, including recent developments of Electron Microscopy (EM) based methods, which are particularly suited for large macromolecular assemblies (4). Still, compared to high-throughput methods for screening protein-protein interactions (i.e. yeast two-hybrid, affinity purification, phage-display etc.), the rate at which novel complex structures are determined experimentally is considerably lower. This uneven growth calls for computational methods to fill the gap.

Several approaches have been developed to address the computational prediction of protein-protein interactions (5). Co-evolution methods, based on correlated amino acid mutations in deep multiple sequence alignments (MSA), are efficiently used to identify interacting proteins based on sequence information alone (6,7). When the 3D structures of the binding partners are available, or can be reliably modelled, docking methods can be used to obtain a three-dimensional model of the complex based on geometric and physicochemical complementarity of the interacting molecules (8–11). Efficiently handling protein flexibility is still one of the major challenges in the development of effective docking simulation software; hence these methods are generally more accurate when little or no conformational change is required for binding. According to the community-wide experiment CAPRI (Critical Assessment of PRedicted Interactions (12)), considerable progress has been made in the field with the development of hybrid modelling strategies, that are able to incorporate available experimental information on the interaction (i.e. crosslinks, NMR, SAXS etc.) as constraints in the simulation of the docking process (13–15). Results from latest assessments show that significantly improved quality of models is obtained when multi-chain template information is available and used for modelling (16).
With more experimentally determined structures of protein complexes becoming available, it has been observed that interacting interfaces are often conserved among homologous complexes (17) and that templates are available for most of the known protein-protein interactions (18). These observations provided the rationale for comparative, or homology modelling, of protein complexes. Similar to comparative modelling of monomeric proteins, the information of a protein's quaternary structure is transferred by homology to another one, and a model of the complex is obtained based on the structures of the interacting homologs, i.e. interologs, as templates (19–21). The approach can be scaled to entire genomes and applied to binary as well as to higher-order protein assemblies (17,18,22,23). As highlighted by the introduction of the first assessment of protein assemblies in the recent CASP XII experiment (24), comparative modelling of protein complexes is receiving much attention and is expected to play a relevant role in the elucidation of the protein quaternary structure space.

SWISS-MODEL (https://swissmodel.expasy.org) was the first fully automated protein homology modelling server and has been continuously improved during the last 25 years (25–30). Its modelling functionality has been recently extended to include the modelling of homo- and heteromeric complexes, given the amino acid sequences of the interacting partners as starting point. Other recently introduced features include the development of a new modelling engine, ProMod3, with increased accuracy of the produced models, and an improved local model quality estimation method (QMEANDisCo) based on a novel version of QMEAN (31).

SWISS-MODEL currently generates ~3000 models a day (~2 models per minute), up from ~1500 models a day in 2014 (30), making it one of the most widely used structure modelling servers worldwide. Its performance is continuously evaluated and compared with other state-of-the art servers in the field. To this aim, we are actively participating to the CAMEO project (Continuous Automated Model Evaluation, https://cameo3d.org) (32), a fully automated blind prediction assessment based on weekly pre-release of sequences from the PDB (33), allowing us to constantly monitor and improve the performance of the server.

MATERIALS AND METHODS

The modelling workflow

In comparative modelling, a 3D protein model of a target sequence is generated by extrapolating experimental information from an evolutionary related protein structure that serves as a template. In SWISS-MODEL, the default modelling workflow consists of the following main steps:

1. **Input data**: The target protein can be provided as amino acid sequence, either in FASTA, Clustal format or as a plain text. Alternatively, a UniProtKB accession code (34) can be specified. If the target protein is heteromeric, i.e. it consists of different protein chains as subunits, amino acid sequences or UniProtKB accession codes must be specified for each subunit.

2. **Template search**: Data provided in step 1 serve as a query to search for evolutionary related protein structures against the SWISS-MODEL template library SMTL (30). SWISS-MODEL performs this task by using two database search methods: BLAST (35,36), which is fast and sufficiently accurate for closely related templates, and HHblits (37), which adds sensitivity in case of remote homology.

3. **Template selection**: When the template search is complete, templates are ranked according to expected quality of the resulting models, as estimated by Global Model Quality Estimate (GMQE) (30) and Quaternary Structure Quality Estimate (QSQE) (23). Top-ranked templates and alignments are compared to verify whether they represent alternative conformational states or cover different regions of the target protein. In this case, multiple templates are selected automatically and different models are built accordingly. To provide the user with the option to use alternative templates than those selected automatically, all templates are shown in a tabular form with a descriptive set of features. In addition, interactive graphical views facilitate the analysis and comparison of available templates in terms of their three-dimensional structures, sequence similarity and quaternary structure features.

4. **Model building**: For each selected template, a 3D protein model is automatically generated by first transferring conserved atom coordinates as defined by the target-template alignment. Residue coordinates corresponding to insertions/deletions in the alignment are generated by loop modelling and a full-atom protein model is obtained by constructing the non-conserved amino acid side chains. SWISS-MODEL relies on the OpenStructure computational structural biology framework (38) and the ProMod3 modelling engine to perform this step. For more detailed information on model building we refer to a dedicated section in Results.

5. **Model quality estimation**: To quantify modelling errors and give estimates on expected model accuracy, SWISS-MODEL relies on the QMEAN scoring function (31). QMEAN uses statistical potentials of mean force to generate global and per residue quality estimates. The local quality estimates are enhanced by pairwise distance constraints that represent ensemble information from all template structures found. For more information on quality estimation we refer to a dedicated section in Results.

SWISS-MODEL allows for further customization of steps 1 and 3. Expert users can directly upload custom target-template sequence alignments, target structures or DeepView project files (26) in separate input forms.

The SWISS-MODEL template library

The SWISS-MODEL Template Library (SMTL), available at https://swissmodel.expasy.org/templates/, is a curated template library, which is updated on a weekly basis according to the new PDB release (33). Every deposited PDB structure is automatically processed, annotated and indexed to support efficient querying of high quality structural data. SMTL entries are organized by quaternary structure assemblies, according to the ‘Biological Assembly’ an-
For visualization of protein structures, users can select to generate a model using SWISS-MODEL is available at https://swissmodel.expasy.org/docs/tutorial.

According to the latest options and features available, an updated version of the documentation is provided to reflect the latest changes of the current SWISS-MODEL 416 structures from PDB with mapping to UniProtKB.

The ProMod3 modelling engine

The modelling engine is the heart of SWISS-MODEL. It builds an atomistic protein model given a template structure and a target-template sequence alignment. Until recently, the software package ProMod-II was used exclusively. ProMod3 has been designed with the aim of providing rapid and flexible prototyping for future modelling developments in SWISS-MODEL.

Like its predecessors, ProMod3 extracts structural information from an aligned template structure in Cartesian space. Insertions and deletions, as defined by the sequence alignment, are resolved by first searching for viable candidate fragments in a structural database. This is a relevant modification, as ProMod-II mainly relied on ab-initio techniques to perform this step. Final candidates are selected using statistical potentials of mean force scoring methods. If no suitable fragments can be found, a conformational space search is performed using Monte Carlo sampling.

A direct comparison between the previous and updated modelling engines has been performed in the context of the CAMEO experiment using 250 target proteins collected during the time range 20 October 2017–13 January 2018. For each target, a template search has been performed using HHblits against the SMTL at the time of the CAMEO submission. The best template, according to the HHblits e-value, and the corresponding target-template sequence alignment served as input for both engines. As shown in Figure 1, models generated with ProMod3 show significantly improved accuracy according to all-atom IDDT (Local Distance Difference Test) score, a superposition-free measure of the deviation of interatomic distances between model and native structures. The same also holds for other commonly used model quality metrics, i.e. GDT-HA (Global Distance Test High Accuracy score) and TM-score (Template Modelling score). Like its predecessors, ProMod3 extracts structural information from an aligned template structure in Cartesian space. Insertions and deletions, as defined by the sequence alignment, are resolved by first searching for viable candidate fragments in a structural database. This is a relevant modification, as ProMod-II main...
score and is then used for the model ranking. To facilitate is updated after model building with the QMEAN global target sequence. To incorporate structural features, GMQE determined protein structures that are homologues to the with ensemble information extracted from experimentally consistency of observed interatomic distances in the model enhance the accuracy of local quality estimates. It assesses the the latest version of QMEAN, QMEANDisCo, further enhances the accuracy of local quality estimates. It assesses the consistency of observed interatomic distances in the model with ensemble information extracted from experimentally determined protein structures that are homologues to the target sequence. To incorporate structural features, GMQE is updated after model building with the QMEAN global score and is then used for the model ranking. To facilitate interpretation of the obtained model quality estimates, the QMEAN global score is transformed to a Z-score, indicating whether the model scores as it would be expected from experimentally determined structures of similar size (31).

Performance comparison with other modelling servers

In order to provide objective assessments of modelling performance, SWISS-MODEL participates in the CAMEO project (https://cameo3d.org) (32). Taking some inspiration from CASP, CAMEO aims to provide a continuous, fully automated, assessment of predictions produced by various modelling servers using a common benchmark dataset of targets. CAMEO target sequences are obtained from the weekly pre-release of new PDB structures and submitted to participating methods at the same point in time. This ensures all servers have access to the same background information, i.e. same structures from PDB or protein sequences in UniProtKB, when running their predictions. Finally, in order to exclude trivial modelling cases, protein sequences exhibiting >85% sequence identity to available PDB structures are not considered in the CAMEO evaluation.

Based on the CAMEO results in the ‘3D Structure Prediction’ category, SWISS-MODEL is consistently ranked among the top-modelling servers for several crucial modelling aspects. Table 1 shows the performance based on a benchmark dataset of 250 targets collected during the CAMEO time range 20 October 2017–13 January 2018. Full data on performance are provided as supplementary materials (Supplementary Tables S1–S7). Notably, SWISS-MODEL has the lowest response time to generate models and excels at model quality for binding sites (IDDT-BS), for high-quality models (IDDT-easy) and for quaternary structure prediction (QS-score). SWISS-MODEL is optimized for comparative protein modelling cases, where high-quality models can be generated and used in a variety of practical research applications (62). For difficult remote homology or de novo modelling targets, other methods perform better in the CAMEO assessment (63–65). It is worth mentioning that among the participating servers, only SWISS-MODEL and Robetta provided results for oligomeric targets. Therefore quaternary structure predictions were assessed on a common subset of oligomeric proteins where both methods returned a model, for a total of 32 targets. Finally, based on the assessment of model confidence, SWISS-MODEL significantly outperforms other modelling servers in providing accurate local confidence estimates of the returned models.

Case study: Modelling the Ferredoxin/Ferredoxin-NADP(+) Reductase complex

To illustrate the new features of SWISS-MODEL, we describe here the modelling of the hetero-dimeric complex formed by Ferredoxin-NADP(+) Reductase (FNR) and its physiological electron donor Ferredoxin (Fd). In higher plants, these proteins are part of the electron transport chain of thylakoid membranes where they catalyse the last step of NADP+ reduction. In non-photosynthetic tissues, i.e. roots, the reaction operates in the opposite direction and is mediated by the tissue specific isoforms of the enzymes (66). Crystal structures of the leaf electron transfer complex FNR:Fd have been reported from Zea mays, provid-
Table 1. Performance comparison in the context of the CAMEO continuous evaluation platform

<table>
<thead>
<tr>
<th>Server</th>
<th>Response time (hh:mm:ss) (N = 168)</th>
<th>IDDT total (N = 168)</th>
<th>IDDT easy (N = 37)</th>
<th>IDDT medium (N = 90)</th>
<th>IDDT hard (N = 41)</th>
<th>IDDT BS (N = 69)</th>
<th>QSScore (N = 32)</th>
<th>Model confidence (N = 168)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWISS-MODEL</td>
<td>00:15:48*</td>
<td>66.22</td>
<td>86.01</td>
<td>69.71</td>
<td>40.67</td>
<td>70.88</td>
<td>63.95</td>
<td>0.85</td>
</tr>
<tr>
<td>HHpredB</td>
<td>01:16:15*</td>
<td>65.95</td>
<td>82.10*</td>
<td>69.68</td>
<td>43.18</td>
<td>71.47</td>
<td>-</td>
<td>0.79*</td>
</tr>
<tr>
<td>NaiveBLAST</td>
<td>01:20:27*</td>
<td>58.93*</td>
<td>82.86*</td>
<td>64.20*</td>
<td>25.76*</td>
<td>63.88*</td>
<td>-</td>
<td>0.68*</td>
</tr>
<tr>
<td>PRIMO</td>
<td>02:12:08*</td>
<td>60.26*</td>
<td>84.51*</td>
<td>65.07*</td>
<td>27.82*</td>
<td>67.30*</td>
<td>-</td>
<td>0.67*</td>
</tr>
<tr>
<td>SPARKS-X</td>
<td>02:35:21*</td>
<td>63.14*</td>
<td>80.06*</td>
<td>65.57*</td>
<td>42.53</td>
<td>67.76*</td>
<td>-</td>
<td>0.54*</td>
</tr>
<tr>
<td>RaptorX</td>
<td>06:28:57*</td>
<td>69.15*</td>
<td>83.35*</td>
<td>72.10*</td>
<td>49.88*</td>
<td>68.85</td>
<td>-</td>
<td>0.65*</td>
</tr>
<tr>
<td>IntFOLD4-TS</td>
<td>32:47:59*</td>
<td>68.41*</td>
<td>83.76*</td>
<td>70.88</td>
<td>49.11*</td>
<td>71.65</td>
<td>-</td>
<td>0.84</td>
</tr>
<tr>
<td>Robetta</td>
<td>37:00:07*</td>
<td>71.60*</td>
<td>85.17</td>
<td>74.00*</td>
<td>54.08*</td>
<td>67.48*</td>
<td>60.20</td>
<td>0.81*</td>
</tr>
</tbody>
</table>

Performance is measured based on a benchmark dataset of 250 targets collected during the CAMEO time range 20 October 2017–13 January 2018. Results from SWISS-MODEL and seven other modelling servers were collected from CAMEO and the performance evaluated on a common subset of targets where all compared servers returned a model. Each column indicates average performance values in terms of Response Time, model accuracy (IDDT, QSScore) and self-assessment of model quality (Model Confidence). IDDT evaluation has further been split according to CAMEOs definition of target difficulty; per column subset sizes are shown in brackets. Asterisks indicate a statistically significant difference (P-value < 0.05) compared to SWISS-MODEL based on paired t-test.

CONCLUSIONS

Computational structural modelling methods have established themselves as a valuable complement to experimen-
Figure 2. Modelling example of the Ferredoxin/Ferredoxin-NADP(+) Reductase heterodimeric complex. (A) Decision tree of templates clustered according to their quaternary structure features: oligomeric state, stoichiometry, topology and interface similarity. Three different clusters are formed based on interface similarity between templates. (B) PPI fingerprint analysis of available template structures. The ratio between interface and surface residue entropy (interface conservation, y-axis) is reported as a function of evolutionary distance (sequence identity, x-axis). Templates corresponding to SMTL ID: 1ewy.1 (in blue) and SMTL ID: 1gaq.1 (in green) show the typical conservation pattern observed for biologically relevant interfaces, with stronger conservation signal in the sequence identity range between 40 and 60%. Considering also remote homologs (below 40% sequence identity), only the interface in template SMTL ID: 1gaq.1 is deemed as conserved (interface/surface conservation ratio below zero). Template corresponding to SMTL ID: 3w5u.1 (in red) displays an interface/surface conservation ratio close to zero, as observed in crystal contacts/artefacts. (C) Structure superposition of available templates. Each template is coloured according to same colouring scheme of Figure 2A and B. Templates corresponding to SMTL ID: 1ewy.1 (in blue) and 1gaq.1 (in green) show similar arrangement of FNR and Fd in the complex. Cross-linked cysteines are shown in sticks. (D) Structure superposition between model and native structure of the root FNR:Fd complex. The model is coloured according to its local quality using a colour gradient from blue (high quality) to red (low quality) as measured by all-atom IDDT score. The native structure of the complex is shown in light gray.

tal structural biology efforts towards increasing our understanding of the protein universe and of its properties. In this endeavour, comparative modelling techniques have matured into fully automated pipelines, providing easy access to reliable 3D models and broadening the spectrum of users and applications of protein models. SWISS-MODEL pioneered the field of fully automated comparative modelling servers 25 years ago and it has been continuously developed and improved since then.

With the new version of SWISS-MODEL presented here, we aimed at extending the scope of automated homology modelling to address the modelling of protein assemblies by efficiently using the information on quaternary structures available in the PDB. The success of this approach clearly depends on the availability of homologous complexes that can be used as templates for modelling. As such, ongoing structural biology efforts leading to structures of macromolecular complexes being determined at unprecedented speed are tremendously beneficial for making our approach increasingly applicable and effective. An important aspect is the ability to handle ambiguous or conflicting information present in available structural data, which is crucial for the development of stable and fully automated pipelines. Here, we showed how our PPI fingerprint analysis and model quality estimates could provide additional criteria to improve the automatic identification of templates, which in turn results into more accurate models and a biologically meaningful representation of their oligomeric state. Finally,
we introduced an improved modelling engine and increased
the precision of model quality estimates, leading to more ac-
curate models and realistic error estimates at the same time.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We would like to thank all SWISS-MODEL users for their
continuous feedbacks and participation to our user surveys,
which motivate and inspire the continuous development of
SWISS-MODEL. We thank sciCORE at the University of
Basel for providing computational resources and system ad-
ministration support.

FUNDING
SIB Swiss Institute of Bioinformatics. Swiss Foundation for
Excellence and Talent in Biomedical Research (to G.S.).
Funding for open access charge: SIB Swiss Institute of
Bioinformatics. ‘Fellowship for Excellence’ international
PhD program of the Biozentrum Basel (to M.B.)
Conflict of interest statement. None declared.

REFERENCES
druggable binding sites at the protein-protein interface. Drug Discov.
Today, 14, 155–161.
2. Nim,S., Jeon,J., Corbi-Verge,C., Seo,M.H., Ivarsson,Y., Moffat,J.,
Tarasova,N. and Kim,P.M. (2016) Pooled screening for antiproli-
in the Protein Data Bank: a status report. Structure, 13, 381–388.
assembly, and evolution of protein complexes. Annu. Rev. Biochem.,
84, 551–575.
Correlated mutations contain information about protein-protein
coevolution gives 3D contacts and structures of protein complexes.
eLife, 3, e00340.
9. Chaudhury,S., Bertrand,M., Schirmer,H., Muruganujan,A. and
Gray,J.J. (2011) Benchmarking and analysis of protein docking
performance in Rosetta v3.2. PLoS One, 6, e22477.
Moreira,J.S., Roel-Tours,J., Melquiond,A.S.J., Geng,C.,
Schaarschmidt,J. et al. (2018) Performance of HADDOCK and a
simple contact-based protein-ligand binding affinity predictor in the
11. Peterson,L.X., Togawa,Y., Esquivel-Rodriguez,J., Terashi,G.,
the assembly order of multimeric heteroprotein complexes. PLoS
the CAPRI experiment. Protein Sci., 14, 278–283.
2075–2081.
Information-driven, ensemble flexible peptide docking using
performance in CAPRI round 37 using LZeRD docking and
template-based modeling with combined scoring functions. Proteins,
86(Suppl. 1), 311–320.
19. Matthews,L.R., Vaglio,P., Rebold,J., Ge,H., Davis,B.P., Garrels,J.,
networks using sequence-based searches for conserved protein-protein
interactions or ‘interologs’. Genome Res., 11, 2120–2126.
relationship between sequence and interaction divergence in proteins.
22. Yu,H., Loscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.D.,
transfer between genomes: protein-protein interologs and
Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Scientific
Rep., 7, 10480.
24. Lafita,A., Bliven,S., Krystkowiavch,A., Bertoni,M.,
Assessment of protein assembly prediction in CASP12. Proteins,
86(Suppl. 1), 247–256.
for automated comparative protein modelling. Biochem. Soc. Trans.,
24, 274–279.
Swiss-PdbViewer: an environment for comparative protein modeling.
Electrophoresis, 18, 2714–2723.
SWISS-MODEL: an automated protein homology-modeling server.
SWISS-MODEL workspace: a web-based environment for protein
modeling with SWISS-MODEL Workspace and the Protein Model
30. Biasini,M., Bienert,S., Waterhouse,A., Arnold,K., Studer,G.,
Schmiidt,T., Kiefer,F., Gallo Cassarino,T., Bertoni,M., Bordoli,L.,
et al. (2014) SWISS-MODEL: modelling protein tertiary and
quaternary structure using evolutionary information. Nucleic Acids
of the absolute quality of individual protein structure models.
Bioinformatics, 27, 343–350.
32. Haas,J., Barbato,A., Behringer,D., Studer,G., Roth,S., Bertoni,M.,
automated model evaluation (CAMEO) complementing the critical
33. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E.,
58, 899–907.
35. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z.,
Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST:


