Protein structure homology modeling using SWISS-MODEL workspace

Lorenza Bordoli, Florian Kiefer, Konstantin Arnold, Pascal Benkert, James Battey & Torsten Schwede

Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50/70, CH 4056 Basel, Switzerland. Correspondence should be addressed to T.S. (torsten.schwede@unibas.ch).

Published online 11 December 2008; corrected online 18 June 2009 (details online); doi:10.1038/nprot.2008.197

Homology modeling aims to build three-dimensional protein structure models using experimentally determined structures of related family members as templates. SWISS-MODEL workspace is an integrated Web-based modeling expert system. For a given target protein, a library of experimental protein structures is searched to identify suitable templates. On the basis of a sequence alignment between the target protein and the template structure, a three-dimensional model for the target protein is generated. Model quality assessment tools are used to estimate the reliability of the resulting models. Homology modeling is currently the most accurate computational method to generate reliable structural models and is routinely used in many biological applications. Typically, the computational effort for a modeling project is less than 2 h. However, this does not include the time required for visualization and interpretation of the model, which may vary depending on personal experience working with protein structures.

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

INTRODUCTION

The three-dimensional structure of a protein provides important information for understanding its biochemical function and interaction properties in molecular detail. However, the number of known protein sequences is much larger than the number of experimentally solved protein structures. As of August 2008, more than 52,500 experimentally determined protein structures were deposited in the Protein Data Bank (PDB)[1]. Yet, this number appears relatively small compared with the more than 6 million protein sequences held in the UniProt knowledge database[2]. Fortunately, the number of different protein fold families occurring in nature appears to be limited[3], and within a protein family, structural similarity between two homologous proteins can be inferred from sequence similarity[4]. Homology modeling (or comparative protein structure modeling) techniques have been developed to build three-dimensional models of a protein (target) from its amino-acid sequence on the basis of an alignment with a similar protein with known structure (template)[5–7]. In cases where no suitable template structure can be identified, *de novo* (a.k.a. *ab initio*) structure prediction methods can be used to generate three-dimensional protein models without relying on a homologus template structure. However, despite recent progress in the field, *de novo* predictions are limited to relatively small proteins and fall short in terms of accuracy compared with comparative models[8–12]. Therefore, homology modeling is the method of choice to build reliable three-dimensional *in silico* models of a protein in all cases where template structures can be identified.

Homology models are widely used in many applications, such as virtual screening, designing site-directed mutagenesis experiments or in rationalizing the effects of sequence variations[13–17]. Stable, reliable and accurate systems for automated homology modeling are therefore required, which are easy to use for both nonspecialists and experts in structural bioinformatics.

Homology modeling

Homology modeling in general consists of four main steps: (i) identifying evolutionarily related proteins with experimentally solved structures that can be used as template(s) for modeling the target protein of interest; (ii) mapping corresponding residues of target sequence and template structure(s) by means of sequence alignment methods and manual adjustment; (iii) building the three-dimensional model on the basis of the alignment; and (iv) evaluating the quality of the resulting model[14,15]. This procedure can be iterated until a satisfactory model is obtained (**Fig. 1**).

Protein structure homology modeling relies on the evolutionary relationship between the target and template proteins. Potential structural templates are identified using a search for homologous proteins in a library of experimentally determined protein structures. From the resulting list of possible candidate structures, a template structure is chosen on the basis of its suitability according to various criteria such as the level of similarity between the query and template sequences, the experimental quality of the solved structures, the presence of ligands or cofactors and so on. Ideally, a large segment of the query sequence should be covered by a single high-quality template, although in many cases, the available template structures will correspond to only one or more distinct structural domains of the protein.

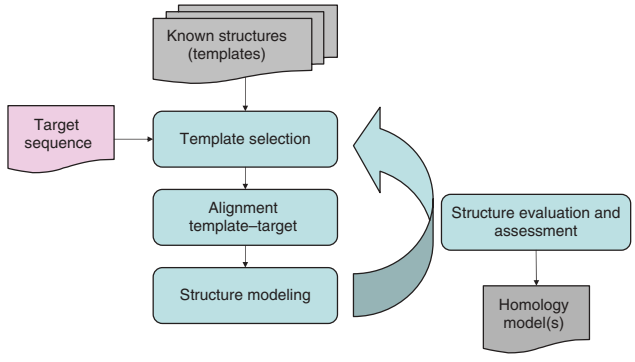

Figure 1 | The four main steps of comparative protein structure modeling: template selection, target–template alignment, model building and model quality evaluation.

Petitioner Merck, Ex. 1012, p. 1

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Estimating the accuracy of a protein structure model is a crucial step in the whole process, as it is the quality of the model that determines its possible applications[18]. The quality of the obtained models will depend on the evolutionary distance between the target and the template proteins. It has been shown that there is a direct correlation between the sequence identity level of a pair of protein structures and the deviation of the Cα atoms of their common core. The more similar two sequences are, the closer the corresponding structures can be expected to be and the larger the fraction of the model that can be directly inferred from the template[4]. As comparative models result from a structural extrapolation guided by a sequence alignment, the percentage of sequence identity between target and template is generally accepted as a reasonable first estimate of the quality of the structurally conserved core of the model. As a rule of thumb, the core Cα atoms of protein models sharing 50% sequence identity with their templates will deviate by ~1.0 Å root mean square deviation from experimentally elucidated structures. Although the atomic coordinates of the three-dimensional model, for regions of the target protein aligned to the template, can be modeled on the basis of the information provided by template structure[5–7], regions that are not aligned with a template (insertions/deletions) require specialized approaches[19–22]. Unaligned regions of the target that are modeled using *de novo* techniques, such as loops, will on average be less accurate than structurally conserved regions of the model on the basis of information derived directly from the template.

As the percentage identity falls below ~30% (in the so-called 'twilight zone'), model quality estimation on the basis of sequence identity becomes unreliable, as the relationship between sequence and structure similarity gets increasingly dispersed[18,23]. With decreasing sequence identity, alignment errors and the incorrect modeling of large insertions become the major source of inaccuracies. Correctly aligning the target sequence with the template is a crucial step and one of the primary sources of errors in the whole modeling procedure. The development of algorithms for sequence comparison and alignment is a major topic of study in bioinformatics, and the advancements in this field have been comprehensively reviewed in ref. 24. For estimating the overall quality of protein structure models and comparing predictions on the basis of alternative alignments[25–27], scoring functions such as statistical potentials of mean force have been developed. Methods that allow identifying local errors in models are currently an active field of research[28–32]. The stereochemical plausibility of the generated models can be assessed using tools such as PROCHECK[33] and WHATCHECK[34], which help to identify amino-acid conformations deviating from expected values for structural features such as bond lengths and angles.

Accuracy and limitations of homology modeling

Comparative modeling relies on establishing an evolutionary relationship between the sequence of the protein of interest and other members of the protein family, whose structures have been solved experimentally by X-ray or NMR. For this reason, the major limitation of this technique is the availability of homologous templates, i.e., only regions of the protein corresponding to an identified template can be modeled accurately. As experimental protein structures are often available only for individual structural domains, it is often not possible to infer the correct relative domain orientation in a model.

Modeling oligomeric proteins, i.e., complexes composed of more than one polypeptide chain, may be straightforward in cases where the complex of interest is similar to a homologous complex of known structure. However, this situation is relatively rare, as most experimental structures in the PDB consist of individual proteins rather than complexes. Modeling complexes from individual components is a daunting task[35] and rarely successful without integrating additional information about the assembly[36].

Comparative protein modeling techniques rely on structural information from the template to derive the structure of the target. Large structural changes, e.g., caused by mutations, insertions, deletions and fusion proteins, are therefore, in general, not expected to be modeled accurately by comparative techniques. Nonetheless, homology models of a protein under investigation can provide a valuable tool for the interpretation of sequence variation and the design of mutagenesis experiment to elucidate the biological function of proteins[16,17,37].

The reliability of different protein modeling methods can be objectively evaluated by examining the quality of predictions made during blinded tests. For example, in CASP7, the 'Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction' in 2006, predictions for 108 homology modeling targets were analyzed in detail to identify progress and limitations of current protein structure prediction methods[11]. Particular emphasis was also given to the analysis of the results of automated prediction servers whose accuracy has significantly increased over the last years. Details about the participating servers and public accessibility are given in Table 1 of ref. 38. Similarly, the EVA[39] project provides a continuous assessment of the stability and accuracy of automated modeling servers on the basis of a large number of blind predictions. SWISS-MODEL was the first comparative modeling server to join the EVA project in May 2000. All results of this evaluation are available at http://eva.compbio.ucsf.edu/~eva/.

Availability

SWISS-MODEL workspace[40] can be freely accessed by the biological community on the Web at http://swissmodel.expasy.org/workspace/. SWISS-MODEL has been the first automated modeling server publicly available[7]. In the meantime, similar services have been developed by other groups, e.g., ModPipe[41], 3D-JIGSAW[42] or M4T[43]. For a more complete listing of other publicly available comparative modeling servers, we refer the readers to the annual Nucleic Acids Research Web server issue[44].

SWISS-MODEL workspace

Each of the four steps in homology modeling requires specialized software as well as access to up-to-date protein sequence and structure databases. The SWISS-MODEL workspace[40] integrates the software required for homology modeling and databases in an easy-to-use, Web-based modeling environment. The workspace assists the user in building and evaluating protein homology models at different levels of complexity—depending on the difficulty of the individual modeling task. A highly automated modeling procedure with a minimum of user intervention is provided for modeling scenarios where highly similar structural templates are available[7,14,45–47]. For more complex modeling tasks where target and template have lower sequence similarity, expert users are given control over the several steps of model building to construct a

Petitioner Merck, Ex. 1012, p. 2

protein model that is optimally adapted to their scientific problem[48]. Modeling can be performed from within a Web browser without the need for downloading, compiling and installing large program packages or databases. The results of different modeling tasks are presented in a graphical summary. As quality evaluation is indispensable for a predictive method like homology modeling, every model is accompanied by several quality checks. In the following sections, we describe the main components of the SWISS-MODEL workspace.

Tools for target sequence feature annotation. Functional and structural domain annotation of the target sequence of interest is the first step toward the identification of a suitable template for building its three-dimensional model. Individual structural domains of multidomain proteins often correspond to units of distinct molecular function[49–51]. Furthermore, the sensitivity of profile-based template detection methods can be enhanced when the search is performed at the domain level rather than searching the whole protein sequence. IprScan, a PERL-based InterProScan[52,53] utility, has been integrated in the SWISS-MODEL workspace for the analysis of the domain architecture of the target protein and the annotation of its functional features. Prediction tools for secondary structure[54], disorder[55] and transmembrane (TM) regions[56] complement the tools for protein sequence analysis and aid the selection of suitable modeling templates for specific regions of the target proteins. In the twilight zone of sequence alignments, applying secondary structure prediction to the protein of interest may help deciding whether a putative template shares essential structural features. Intrinsically unstructured regions in proteins have been associated with numerous important biological cellular functions, from cell signalling to transcriptional regulation[57,58]; several examples of such disordered regions undergoing the transition to an ordered state upon binding their ligand proteins have been reported[59]. Prediction of disordered and trasmembrane regions therefore complement the analysis of protein domain boundaries and functional annotation of the target protein.

Tools for template identification. The SWISS-MODEL workspace provides a set of increasingly sensitive sequence-based search methods for template detection that are applied depending on the evolutionary divergence between the target protein and the closest structurally characterized template protein. Close homologs of the target can be identified using a gapped BLAST[60] query against the SWISS-MODEL Template Library (SMTL)[40]. When no closely related templates are found, or can be identified only for some segments of the target protein, more sensitive approaches for detecting evolutionary relationships are provided. (i) In the iterative profile Blast approach[60], which has been initially introduced as PDB-Blast by Godzik and coworkers, a profile for the target sequence is compiled from homologous sequences by iterative searches of the NR database[61] and used subsequently to search SMTL for homologous structures. (ii) Alternatively, to detect more distantly related template structures, a Hidden Markov Model (HMM) for the target sequence is built on the basis of a multiple sequence alignment, similarly to the profile Blast approach discussed above. The HMM for the target sequence is subsequently used to search against the template library of HMMs generated for a nonredundant set of the sequences of the SMTL template library culled at 70% sequence identity. HMM building, calibration and library searches are performed using the HHSearch (v. 1.5.01) software package[62]. For the selection of a suitable template, the following issues need to be considered:

1. The selection of the best template structure not only depends on sequence similarity but should also take into account other factors, such as experimental quality, bound substrate molecules or different conformational states of the template. For example, certain proteins undergo large conformational changes upon substrate binding as observed, e.g., between the apostructure and ATP- and ADP-bound forms of enzymes in the nucleotide kinase family[63]. Depending on the planned model applications, such as structure-based ligand design, it is necessary to choose a structural template in the correct conformation.
2. For low-homology templates, the InterPro functional annotation of the target sequence can be used to verify that putative templates share essential functional features.
3. If the target–template alignment falls within the 'twilight zone' of sequence alignments (i.e., below 30% sequence identity), secondary structure prediction of the target protein may help to decide whether a putative template shares structural features with your protein and may therefore be used as template.
4. Predicted disorder regions may indicate the boundaries of protein domains and provide additional functional annotation of the protein.

Modeling. *Automated mode*: If the alignment between the target and the template sequences displays a sufficiently high similarity, a fully automated homology modeling approach can be applied. As a rule of thumb, automated sequence alignments are sufficiently reliable when target and template share more than 50% of sequence identity. Submissions in 'Automated mode' require only the amino-acid sequence or the UniProt[2] accession code of the target protein as input data. The hierarchical approach for template detection of the modeling pipeline will automatically select suitable templates on the basis of a Blast search or using an adapted sequence-to-HMM comparison HHSearch protocol[64]. In cases where several similar template structures are available, the automated template selection will favor high-resolution template structures with good-quality assessment. Optionally, a specific template from the SMTL template library can be specified.

Alignment mode: For more distantly related target and template sequences, the number of errors in automated sequence alignments increases[23]. This poses a major problem for automated homology modeling, as current methods are not capable of recovering from an incorrect input alignment. In many molecular biology projects, multiple sequence alignments are often the result of extensive theoretical and experimental exploration of a family of proteins. Such alignments can be used for comparative modeling using the 'Alignment mode' if at least one of the member sequences represents a protein for which the three-dimensional structure is known. The 'Alignment mode' allows the user to test several alternative alignments and evaluate the quality of the resulting models to achieve an optimal result.

Project mode: In the so-called 'twilight zone' of sequence alignments, when the sequence identity between target and template is below 30%, it is advisable to visually inspect and manually edit the

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Petitioner Merck, Ex. 1012, p. 3

target–template alignment. This will lead to a significant improvement of the quality of the resulting model. The program DeepView (Swiss-PdbViewer)[48] can be used to display, analyze and manipulate modeling projects. DeepView project files contain one or more superposed template structures and the alignment between the target and template(s). Project files are also generated by the workspace template selection tools and are the default output format of the modeling pipeline. Project files with modified alignments can then be saved to disk and submitted as 'Project mode' to the workspace for model building by the SWISS-MODEL pipeline, thereby giving the user full control over essential modeling parameters: several template structures can be compared simultaneously to identify structurally conserved and variable regions and select the most suitable template. The placement of insertions and deletions in the target–template alignment can be visualized in their structural context and adjusted accordingly.

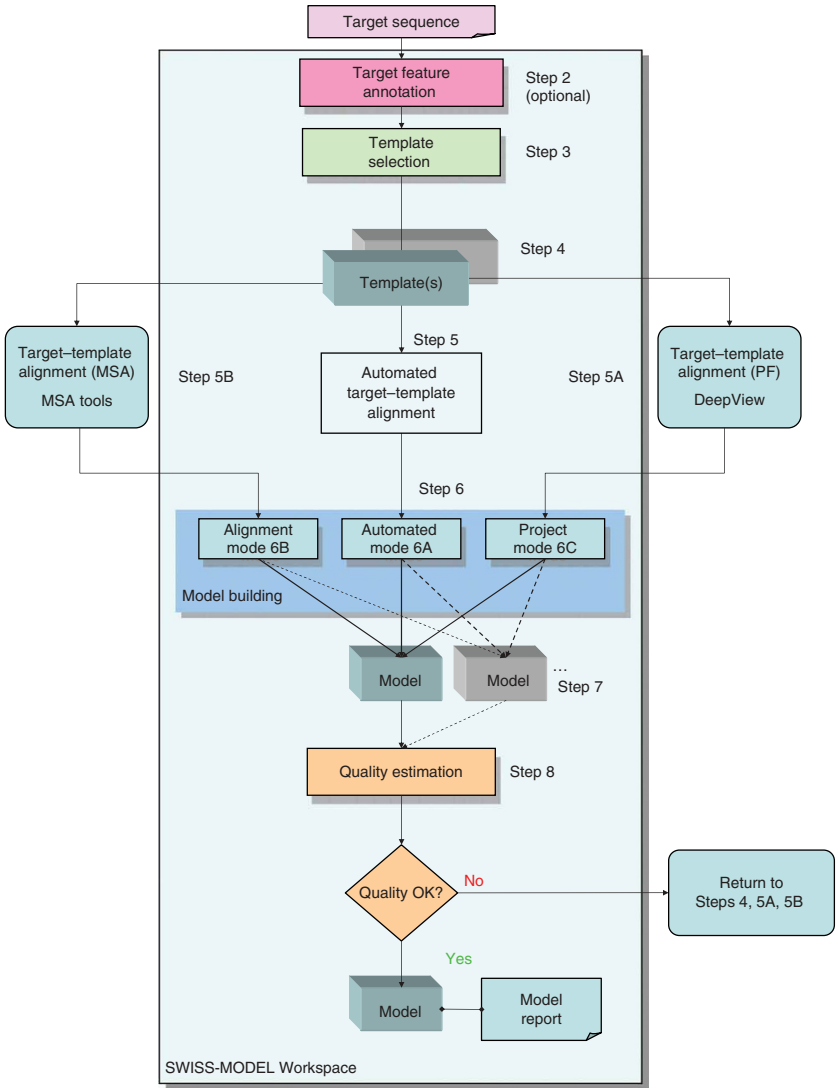

Figure 2 | Workflow of comparative protein structure modeling using SWISS-MODEL workspace. Starting from the amino-acid sequence of a 'Target' protein, three alternative routes for model building are provided—depending on the difficulty of the modeling task. Individual steps are described in detail in PROCEDURE.

Protein structure assessment and model quality estimation. The percentage of sequence identity between target and template is generally accepted as a reasonable first estimate of the quality of a model. However, the accuracy of individual models may vary significantly from the expected average quality due to suboptimal target–template alignments, low template quality, structural flexibility or inaccuracies introduced by the modeling program. Individual assessment of each model is therefore essential. As a global indicator of the quality of a given model, the results of QMEAN[65], a composite scoring function for model quality estimation, and DFIRE[30], an all-atom distance-dependent statistical potential, are provided in the SWISS-MODEL workspace. However, a good global score does not guarantee that important functional sites of a protein have been modeled correctly. Therefore, tools for local model quality estimates are included: graphical plots of ANOLEA mean force potential[28], GROMOS empirical force field energy[66] and the neural network-based approach ProQres[32] are provided as indicators for local model quality. Finally, Whatcheck[34] and Procheck[33] reports enable the user to assess the conformational quality of both models and template structures.

In this protocol, we describe in a step-by-step procedure (**Fig. 2**) how users can benefit from the integrated design of the SWISS-MODEL workspace to build and assess the accuracy of homology models.

MATERIALS

EQUIPMENT

- Amino-acid sequence of the protein to be modeled
- A computer with access to the Internet and a Web browser
- A multiple protein sequence alignment, including at least the sequences of the target protein and the template structure (optional; see Step 6B in PROCEDURE for information on sequence alignment formats)
- DeepView for protein structure analysis and visualization (optional software). DeepView can be freely downloaded from the ExPASy website (http://www.expasy.org/spdbv/)

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Petitioner Merck, Ex. 1012, p. 4

PROCEDURE

Access and personal user account

1| Access and create a personal account for the SWISS-MODEL workspace at http://swissmodel.expasy.org/workspace/. The user data are stored in a password-protected personal user space, which is identified by the user's email address. It is also possible to access the workspace system anonymously without providing an email address. However, it is then necessary to bookmark the URLs of individual work units in the Web browser to be able to retrieve the results once the browser session has been closed. Once logged in your personal user account, individual modeling tasks are organized in work units under 'Workspace'; their current computational status is represented graphically (**Fig. 3**).

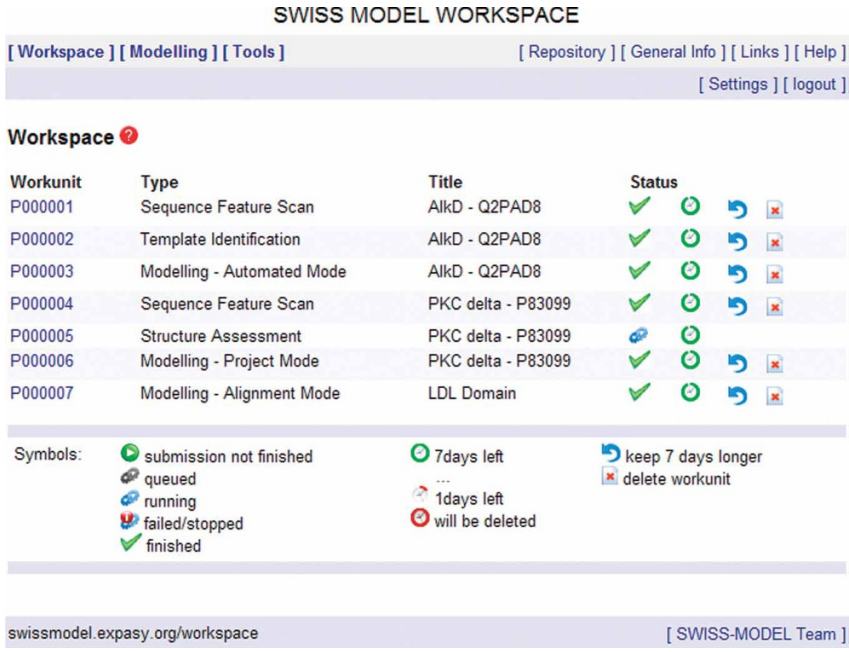

Figure 3 | Example of a personal user workspace. In SWISS-MODEL workspace, individual modeling tasks are organized in work units; their current computational status is represented graphically.

Sequence feature annotation

2| Examine your target sequence. The results of this analysis will assist you in deciding which of the possible template(s) (obtained in Step 3) to use to build homology model(s). Submit your protein sequence (as plain text, in FASTA format or its UniProt Accession Code) to one or more of the tools available in the 'Sequence Features Scan' session of the server, you find under 'Tools': use option A for InterPro domain scan, option B for PsiPred, option C for DISOPRED and option D for MEMSAT.

(A) InterPro domain scan

(i) InterPro domain scan[52] identifies known protein domains and functional sites of the target sequence and possibly assigns the protein to a specific family. The following databases, currently part of the InterPro scan method, can be selected: HMMPfam—the target sequence is searched against the Pfam[67] database, a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families; ProfileScan—the target sequence is searched against the profiles collection of PROSITE[68], a database of protein families and domains, and it consists of biologically significant sites, patterns and profiles that help in identifying to which known protein family a sequence belongs; ScanRegExp—the target sequence is scanned for biologically significant patterns contained in the PROSITE database collection, e.g., enzyme catalytic sites, phosphorylation sites and so on.

(ii) The occurrence of domains and functional sites are displayed on the target sequence. Domain boundaries and links to InterPro database instruct about distinctive features of a given functional domain or provide documentation relative to a specific protein family.

(B) PsiPred

(i) PsiPred[54] predicts secondary structure elements of the target sequence. The graphical representation shows the probability of a given residue of being part of an alpha helix (H), extended beta strand (E) or a coil region (C).

(C) DISOPRED

(i) DISOPRED2[55] predicts the occurrence of disordered regions in the target protein. The probability of being disordered (ranging from 0 to 1) is plotted for each position in the sequence. The 'output' and 'filter' curves represent the raw and filtered scores from the linear SVM classifier (DISOPREDsvm), respectively. Both outputs from DISOPREDsvm are included to allow the user to identify shorter, low-confidence predictions of disorder. Asterisks (*) and dots (.) denote predicted disorder and order, respectively. DISOPRED2 predictions are given at a default false-positive rate threshold of 2%, but this value can be changed by the user.

? TROUBLESHOOTING

(D) MEMSAT

(i) MEMSAT[56] predicts the occurrence of putative TM segment in the protein. Central TM helix segments are indicated with 'X' in the output sequence. Information about the predicted TM topology is also provided.

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Petitioner Merck, Ex. 1012, p. 5

Template identification and target template alignment

3| Submit your target sequence (as plain text, in FASTA format or its UniProt Accession Code) to one or more of the template identification tools of the 'Template Identification' session you find under 'Tools'. The server provides access to a set of increasingly complex and computationally demanding methods: use option A for BLAST, option B for PSI-BLAST and option C for HMM-HMM-based searching.

(A) BLAST

(i) Closely related homologous templates are identified by running a gapped BLAST search[60]. Adjust standard BLAST parameter like *E*-value cutoff or the choice of the substitution matrix to alter the sensitivity or specificity of your search.

? TROUBLESHOOTING

(B) PSI-BLAST

(i) More divergent template structures can be identified using iterative, profile-based BLAST[60].

(ii) Profile generation: selectivity and sensitivity of the search can be adjusted in the profile generation step by altering the number of iterations and the inclusion threshold for building the target profile. A more permissive *E*-value threshold and a greater number of iterations will increase the sensitivity of your search. Note that the inclusion of false positives, i.e., proteins that do not belong to the family of interest, during profile building can cause a drift in the search and lead to an increased false-positive rate among your hits.

(iii) *Profile search*: in the template library search step, the balance between selectivity and sensitivity can be adjusted by the choice of the substitution matrix.

? TROUBLESHOOTING

(C) HMM-HMM-based searching

(i) Distantly related templates can be identified using HMM-based profile matching using HHSearch[62]. A profile of the query sequence is generated and used for identifying matching HMM profiles in the template library. As this approach is computationally more intensive, compared with methods described in Steps 3A and B, the query is performed against a reduced version of the PDB database (culled at the 70% sequence identity level).

4| Select one or more structures from the result hit list as template to build comparative models. Results of template selection (Step 3) and domain identification (Step 2A) are displayed in a condensed graphical overview. This combined view allows you to analyze template coverage with respect to the domain boundaries and to identify templates spanning one or more domains of the target. Bars indicating matching regions will link you to the underlying target–template alignment and links to the SMTL library are available to facilitate the choice of a suitable template.

? TROUBLESHOOTING

5| Once you have selected one or more suitable templates, the following options are possible to improve the initial target–template alignment: Option A—DeepView Project or option B—alternative sequence alignment methods.

▲ CRITICAL STEP This is a particularly critical step, as homology modeling techniques cannot recover from an incorrect starting target–template alignment.

(A) DeepView Project

(i) The target–template sequence alignments generated by the different template database search techniques can be used as the basis for the subsequent model creation. The alignments can be downloaded as DeepView project file, which contains the target sequence aligned to the template structure.

(ii) The program DeepView allows you to display and analyze the alignment in the structural context of the template to manually adjust misaligned regions.

? TROUBLESHOOTING

(iii) Once you have finished editing the alignment, save the project file on the local disk and submit it to the 'Project Mode' of the Modeling session for model building (Step 6C).

(B) Alternative sequence alignment methods

(i) You might also want to apply alternative sequence alignment methods by using multiple sequence alignment programs to align the target and the template sequences obtained in Step 3. For a list of the most widely used sequence alignments tools, please refer to ref. 24, Table 1 therein.

(ii) The obtained sequence alignment between target and template (and additional homologous proteins) can be submitted to the 'Alignment mode' of the modeling session for model building (Step 6B).

Modeling

6| To obtain an homology model of your target sequence, you can choose among three different approaches—accessible through the 'Modeling' session of the server—whose applicability depends primarily on how distantly related your protein and the homologous template are: option A—automated mode; option B—alignment mode; or option C—project mode.

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Petitioner Merck, Ex. 1012, p. 6

(A) Automated mode

(i) In cases where the target–template similarity is sufficiently high to allow for unambiguous sequence alignment, homology modeling can be fully automated. Submit your target sequence as plain text, FASTA format or its UniProt accession code.

(ii) Optionally, a specific template, e.g., identified by a Blast search in Step 3, can be specified by its PDB identifier and chain ID. Make sure that the specified template ID is present in the SWISS-MODEL template library.

? TROUBLESHOOTING

```
>TARGET
QGQEPPPEPRITLTVGGQPVTFLVDTGAQH
SVLTQNPGPLSDRSAWVQGATGGKRYRWTT
DRKVHLATGKVTHSFLHVPDCPYPLLGRDL
LTKLKAQI;
QGQEPPPEPRITLTVGGQPVTFLVDTGAQH
SVLTQNPGPLSDRSAWVQGATGGKRYRWTT
DRKVHLATGKVTHSFLHVPDCPYPLLGRDL
LTKLKAQI
```

Figure 4 | Example of the input format for an oligomeric target sequence (**Box 1**).

(B) Alignment mode

(i) With decreasing sequence similarity between target and template, the number of errors in automatically generated sequence alignments increases. Therefore, you might choose to submit an alignment generated by alternative sequence alignment tools (Step 5B). Provide a pairwise or multiple sequence alignment as input alignment in FASTA, MSF, ClustalW, PFAM or SELEX format.

? TROUBLESHOOTING

(ii) After the alignment has been converted into a standard format, indicate which sequence corresponds to the target protein and which corresponds to a protein with known structure in the template library.

? TROUBLESHOOTING

(iii) Submit your alignment for model calculation. Note that the SWISS-MODEL pipeline used for the modeling process might introduce minor heuristic modifications to improve the placement of insertions and deletions during model building.

(C) Project mode

(i) In the twilight zone of sequence alignments, visual inspection and manual manipulation of the target–template alignment can significantly improve the quality of the resulting model. Using Project mode, you can submit Project files that you have obtained in Step 5A after adjusting the alignment in DeepView.

? TROUBLESHOOTING

(ii) In project mode, you can also submit projects generated directly inside DeepView. With this option, it is possible to generate models using templates that are not part or not yet present in the SMTL library.

(iii) *Oligomer modeling*: template-based modeling of oligomeric assemblies (**Fig. 4**) is possible using DeepView and subsequently submitting the file to the Project Mode (**Box 1**).

7| After completion of the modeling procedure, the results are stored in the workspace and, if specified in your personal setting, you will be notified of the completion. Coordinates of the model, the underlying alignment, log files and quality evaluations can be accessed and downloaded from the personal workspace (**Fig. 5**). The model coordinates are available in PDB or DeepView project file format. The latter allows you to further inspect and manually modify the target–template alignment. Modified project files can then be saved to disk and submitted as 'Project mode' to the workspace for a further iteration of the model-building cycle (Step 6C). Energy profiles from the ANOLEA statistical potential[28] as well as the GROMOS force field[66] are

BOX 1 | OLIGOMER MODELING

(i) Determine the correct quaternary state of the template. Asymmetric units of PDB files often do not correspond to the correct biological assembly of a protein. Assembled coordinate files of the most likely biological assembly of the template can be retrieved from PQS[77], PISA[78] or the PDB.

! CAUTION Homology between two proteins is not necessarily sufficient to signify that they share the same quaternary structure.

(ii) Download and save the oligomer template coordinates as PDB file to your local disk.

(iii) Open the file in DeepView and remove all nonamino-acid groups, such as ions, ligands, OXT and so on from the template (unless they are at the very end of the file). You can do this by selecting the groups in the control panel of DeepView and remove the selected residues ('Build' menu).

(iv) Make sure each chain (protomer) has a unique chain identifier, e.g., 'A', 'B' and so on. Coloring the molecule by chain helps to check. You can rename chains with DeepView ('Edit → Rename' menu).

(v) Create a FASTA file with the target sequences for each chain, i.e., 'A', then 'B' and so on, separated by semicolons. For hetero-oligomers, make sure the order is the same as in the template (**Fig. 4**).

(vi) Adjust target–template alignment in DeepView. Load the FASTA file into DeepView ('SwissModel' Menu) and generate a preliminary target–template alignment ('Fit → Fit raw sequence' Menu). Open the alignment window and adjust the alignment. Be sure not to align residues of different chains or to align amino-acid residues to ligand (HETATM) groups or the C-terminal oxygen group (OXT) in the template. Make sure all insertions and deletions are correctly positioned in the structural context.

(vii) Save the project to your local disk and submit the file to the Project mode of SWISS-MODEL workspace for model building (Step 6C).

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Petitioner Merck, Ex. 1012, p. 7

calculated in the course of the modeling procedure, and the corresponding plots are also accessible from the results page. The percentage sequence identity between target and template on the basis of the alignment used to build the model is also reported on the results page.

? TROUBLESHOOTING

Quality estimation

8| To estimate the quality of your model(s), submit it to the programs provided in the 'Structure Assessment' section under 'Tools,' using the following options: option A for sequence identity; option B for stereochemistry check; option C for global model quality estimation; and option D for local model quality estimation. Some of the tools described below can help identify incorrect regions in the predicted protein structure. One possibility to cope with the uncertainties in comparative modeling (especially when the sequence identity between target and template is low) is to build multiple models on the basis of alternative templates and/or alignments (Step 3–6) and to subsequently select the most favorable model.

▲ CRITICAL STEP Protein structure models generated by comparative modeling may contain errors and thus need to be treated with caution. Often, the quality varies between parts of the model.

? TROUBLESHOOTING

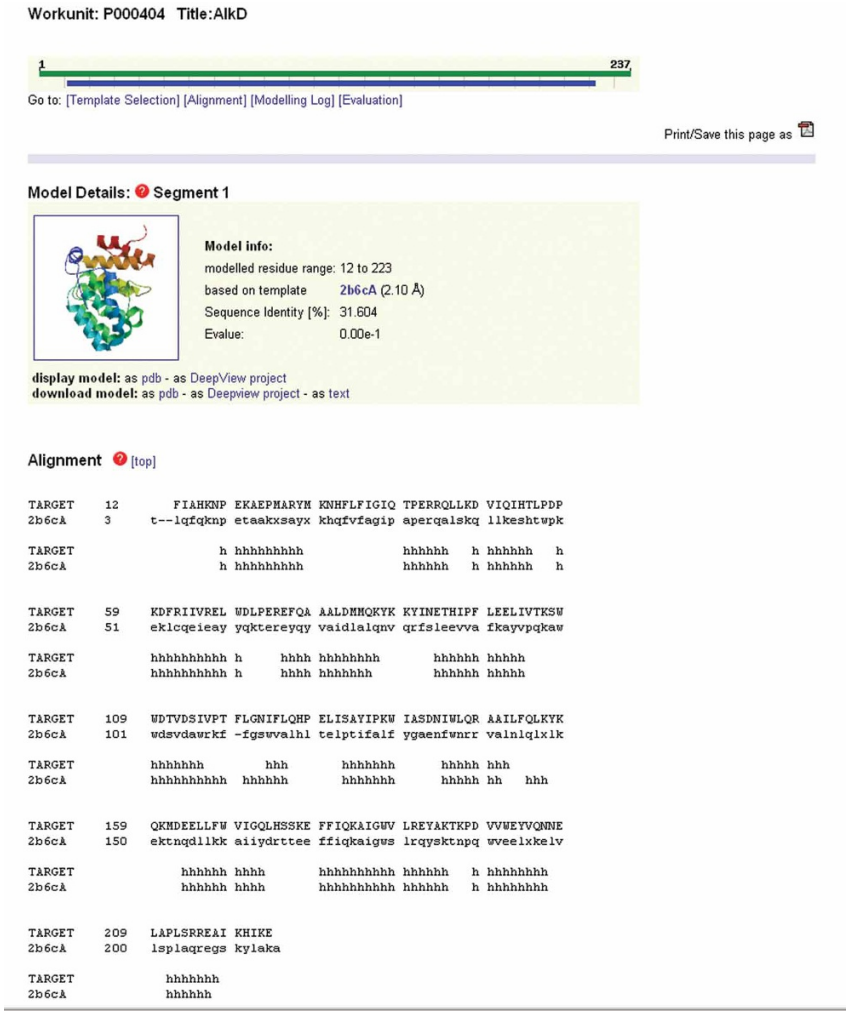

Figure 5 | Typical view of a SWISS-MODEL workspace result. In this example, a model for methylpurine-DNA glycosylase has been generated in automated mode. Upper panel: the green line represents the target sequence (237 residues). Blue lines indicate for which segments of the target models have been generated, in this case for residues 12–223. Middle panel: information on the template structure (2b6c, chain A) and quality (sequence identity, *E*-value) of the target–template sequence alignment shown in the lower panel. Model coordinates can be displayed within the Web browser window by clicking on the preview image or downloaded for manipulation with external software.

(A) Sequence identity

(i) The percentage sequence identity between target and template is a good predictor of the accuracy of a model. Model accuracy steadily increases with increasing sequence identity.

(B) Stereochemistry check

(i) The stereochemical plausibility of the model can be analyzed with the tools WHATCHECK[34] and PROCHECK[33]. Deviations from ideal stereochemical values are reported by these programs.

(C) Global model quality estimation

(i) The DFIRE statistical potential[30] as well as the QMEAN composite scoring function[65] both return a pseudo-energy for the entire model.

(D) Local model quality estimation

(i) The following tools are available for analyzing the local (per-residue) model reliability that can help in identifying potentially incorrect regions in the model: ProQres[32]—an artificial neural network trained to predict the local model quality on the basis of the analysis of atom–atom contacts, residue–residue contacts, solvent accessibility surfaces and secondary structure propensities. The plot shows the local reliability of the model ranging from 0 (unreliable) to 1 (reliable) for each residue in the sequence; ANOLEA[28] is a statistical potential that can be used to analyze the packing quality of the model on the basis of nonlocal atomic interactions. The plot shows the ANOLEA pseudo-energy for each amino acid in the sequence. Negative values (colored in green) indicate that the amino acid is in a favorable environment, whereas positive values (colored in red) suggest that this part of the model has been incorrectly built; the GROMOS[66]

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Petitioner Merck, Ex. 1012, p. 8

empirical force field is used to calculate the energy of each residue in the model. The graphical representation shows position in the sequence against the empirical force field energy. Negative values (colored in green) represent energetically favorable conformations, whereas positive values (colored in red) indicate unfavorable conformations.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Solution
2B and C	Secondary structure prediction predicts a strand helix and the disorder prediction predicts the same region to be disordered	Examples of disordered regions undergoing the transition to an ordered state upon binding their ligand proteins have been reported[59]. In case the predicted region aligns to a known template structure, check if the template has been solved in complex with a binding partner or is otherwise known to undergo structural rearrangement
3A	BLAST reports too many matches	Change the *E*-value cutoff for reporting hits to force BLAST to report only hits with a low *E*-value
	BLAST does not report any results	When no suitable templates are identified, or only parts of the target sequence are covered, two approaches for more sensitive detection of distant relationships among protein families are provided (3B and 3C)
3A and B	How will the choice of the substitution matrix influence the output of Blast/Profile Blast?	Use a substitution matrix adapted to the expected divergence of the searched sequences. For the BLOSUM family of matrices, the higher the matrix index is (i.e., BLOSUM 80), the more selective your search will be: it will exclude false positives but possibly miss true positives (closest to PAM120). Vice versa, the lower the index is (i.e., BLOSUM 45), the more sensitive the search will be: more true matches will be identified, but eventually, more false positives will be included (closest to PAM250)
3B	Profile Blast report too many matches	Change the *E*-value cutoff for reporting hits, or in the template library search step choose a library where the sequences of the templates are clustered at a lower sequence identity, e.g., ExPDB 70
4	The template identification methods cover only part of the sequence of my protein	The sensitivity of profile-based template detection methods (Step 3B and C) can be increased when the search is performed at the domain level rather than using the whole target sequence
	The template identification method predicts two templates with different structures	*Similar structures in different conformation:* protein structures can undergo large conformational changes, e.g., upon binding of ligand or post-translational modification. From the list of possible template structures, select the one most suitable for your application on the basis of the annotation provided, e.g., presence of ligands or cofactors and so on.
		Template structures with different folds: ambiguous results in fold assignment are expected if the evolutionary relationship between the target and possible template structures is too weak to be reliably detected by sequence-based methods, or no related template structure has been solved. Template structures with unclear evolutionary relationship to the target should not be used for homology modeling, unless supported by additional (experimental) evidence
	No templates are found for the protein or domain of interest	In this case, it is not possible to produce a reliable three-dimensional model for the protein by homology modeling. Alternatively, one can attempt to apply *de novo* prediction methods. However, the results of these types of prediction are currently far less accurate than comparative techniques and often not sufficient for specific biological applications[10]
	Can different templates, covering different regions of my protein, be combined to obtain a comparative model for the full length of the protein?	Nonoverlapping templates cannot be combined, as the relative orientation of the different structures is unknown. It is, however, in principle possible, if the different templates are significantly overlapping (e.g., more than 20–40 amino acids). This feature is currently not supported by SWISS-MODEL workspace; users are referred to other modeling programs supporting this option, such as Modeller[6]

(continued)

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Petitioner Merck, Ex. 1012, p. 9

TABLE 1 | Troubleshooting table (continued).

Step	Problem	Solution
5A and 6C	Where can I find information on how to use the program DeepView?	Manuals for DeepView can be found on the program website: http://www.expasy.org/spdbv/. It is highly recommended to start by following the tutorial provided by Gale Rhodes (University of Maine): http://www.usm.maine.edu/~rhodes/SPVTut/
6A and 6B	The template of interest is not present in the SMTL library	The SMTL library is updated biweekly, i.e., it might take a few days until newly released PDB structures are included. You can check if a specific PDB entry is available by querying the SMTL library in the 'Tools' section of the server. Alternatively, you can use DeepView to create a model project file on the basis of any template structure independently if this structure is part of PDB or SMTL
6B	Multiple sequence alignment is not correctly recognized by the server	Please make sure the alignment is in one of the supported formats. Use short unique names for sequences in the alignment; avoid nonalphanumerical characters. Good examples: 'THN_DENCL', 'P01542', '1crnA' and so on
7	Unable to obtain a model from the server	In the majority of cases this is due to a poor alignment between the target and template sequences. Take time to carefully edit the alignment as suggested in Step 6A and B
	Despite careful checking of the target–template alignment, the server does not successfully deliver a model	This is usually the case if the target–template alignment contains large gaps (insertions and deletions) that were not successfully reconstructed. If this is the case, consider using other modeling programs, e.g., Modeller[6]. However, keep in mind that the results of *de novo* loop modeling techniques are less reliable than the template-based part of the model when using the model to answer the biological questions of interest
8	How to proceed when incorrect regions are identified and how to interpret them	There are several possible explanations for regions predicted as potentially incorrect (i.e., having low ProQres scores and/or high ANOLEA energies), e.g., alignment errors, incorrect modeling of insertions, unfavorable side-chain packing or false-positive assignment by the program itself. The identified regions of low reliability should be further analyzed by visually inspecting the alignment and the model. A model should always be interpreted with respect to its future application: a model with local errors outside the region of interest, such as the active site, can nevertheless be valuable for certain experiments. On the contrary, if, e.g., surface loops contain residues known to be involved in function, one needs to proceed with great caution when using the model for refining functional hypothesis

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

ANTICIPATED RESULTS

As an example, we apply the protocol described here to model the bacterial methylpurine-DNA glycosylase (AlkD, Uniprot AC: Q2PAD8) and the *Drosophila* putative protein kinase C delta type (Pkcdelta, UniProt AC: P83099). Please note that the results presented here illustrate a representative example at the time of writing. As sequence and structure databases are continuously updated, new template structures may become available at a later point and may lead to different, in general, better, modeling results.

AlkD is a DNA gylcosylase[69] that functions as a DNA alkylation repair enzyme[70]. AlkD belongs to a newly characterized DNA glycosylase superfamily[71], for which no structures have yet been solved experimentally. Domain annotation (Step 2A) indicates that the protein belongs to a multihelical fold called the armadillo-like fold[72]. This is in agreement with the results of the secondary structure prediction analysis (Step 2B), predicting almost exclusively alpha-helices. A search for suitable structure templates (Step 3) yields a highly significant match spanning almost the entire length of the target protein, a putative DNA alkylation repair enzyme from *Enterococcus faecalis* (PDB: 2B6C) solved by the Midwest Center for Structural Genomics. This template was used by Dalhus *et al.*[73] to built a comparative model to elucidate the mechanism of AlkD.

The BLAST alignment between the target and the template displays only a single gap and a sufficiently high level of sequence similarity for it to be modeled using the automated mode of SWISS-MODEL workspace (Step 6A, **Fig. 5**). The location of the single gap in the target and sequence alignment can further be investigated in the structural context with the help of DeepView. The project file resulting from the automated step containing the template and the modeled protein can be opened in DeepView and the target–template alignment can be visualized with the help of the Alignment Window of DeepView. Secondary structure elements of the template can be highlighted in different colors using the color menu of the software. In the alignment resulting

Petitioner Merck, Ex. 1012, p. 10

from the automated mode, there is an inserted amino acid in the target sequence in a position that corresponds to an internal alpha helix residue of the template. We assume that the position of the gap could be improved by shifting it to the loop region connecting two adjacent alpha helices. The alignment can be edited directly within the alignment window of DeepView (Step 5A). The resulting modified project file is then saved locally and submitted to the Project Mode of the server for model building (Step 6C).

Dalhus *et al.*[73] predicted the location of a putative binding pocket in the model by residues conservation analysis of homologous proteins. The binding site of the obtained model was then used to design site-specific mutations to characterize the role of specific residues in the catalysis of DNA repair. Further insights into the mechanism of activity of the enzyme were gained by combining the obtained model with DNA coordinates extracted from the homologous protein AlkA.

In the second example, we build a model for a putative protein kinase C delta from *Drosophila* (Pkcdelta δ, UniProt AC: P83099). Domain annotation (Step 2A, **Fig. 6**) confirms that the target belongs to the protein kinase C (PKC) family[74]. The PKC family members can be grouped into three classes: (i) the conventional PKCs (α, γ and β1 and β2), which requires diacylgycerol (DAG), phosphatydilserine (PS) and calcium for activation; (ii) the novel PKCs (δ, ε, η/λ, θ), which are activated by DAG and PS but are insensitive to calcium; and (iii) the atypical PKCs (ζ and ι/λ), which require only PS for full activity[75].

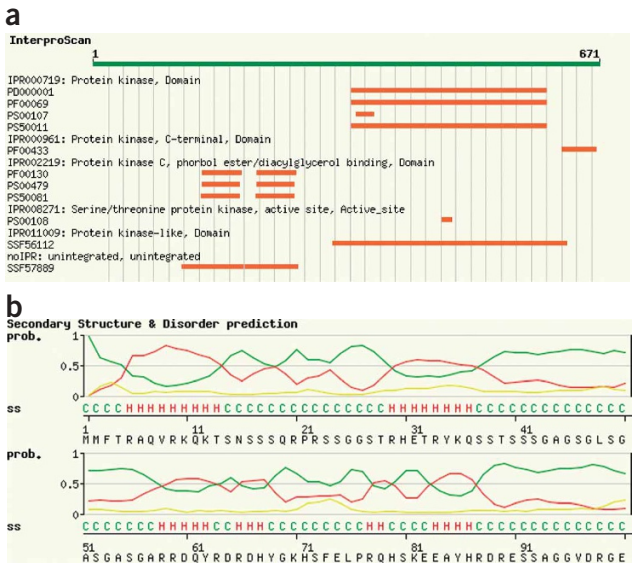

Figure 6 | Target sequence annotation for the putative protein kinase C delta from *Drosophila* (PKC δ, UniProt AC: P83099). (**a**) Three functional domains are indentified using InterPro scan: two PE/DAG (Phorbol esters/Diacylglycerol)-binding domains and a PKC domain. (**b**) Secondary structure prediction for the N-terminal 100 residues of the target sequence.

Several statistically significant matches for suitable templates are reported for the protein kinase domain (Step 4). Each template in the list is linked to the SMTL, and from there to other external resources, to allow for verification and a plausibility check. We have selected template 2JED (chain A) to build the model for our protein. Template 2JED corresponds to the crystal structure of the human kinase domain of the PKC θ, which belongs to the same class as our putative PKC δ. As in the previous example, the target–template alignment (derived from the iterative profile Blast search) is inspected with the help of DeepView, e.g., to verify that the residues corresponding to the typical signatures for the serine/threonine protein kinase active site (Prosite Accession number PS00108) and for the ATP-binding motif (Prosite Accession number PS00107) are correctly mapped in the target and template alignment. This can be done easily with the 'search for PROSITE pattern' function in the 'Edit' menu of DeepView. Subsequently, the target–template alignment is saved as project file and submitted to the Project Mode to obtain a model for the protein kinase domain of the protein (**Fig. 7**).

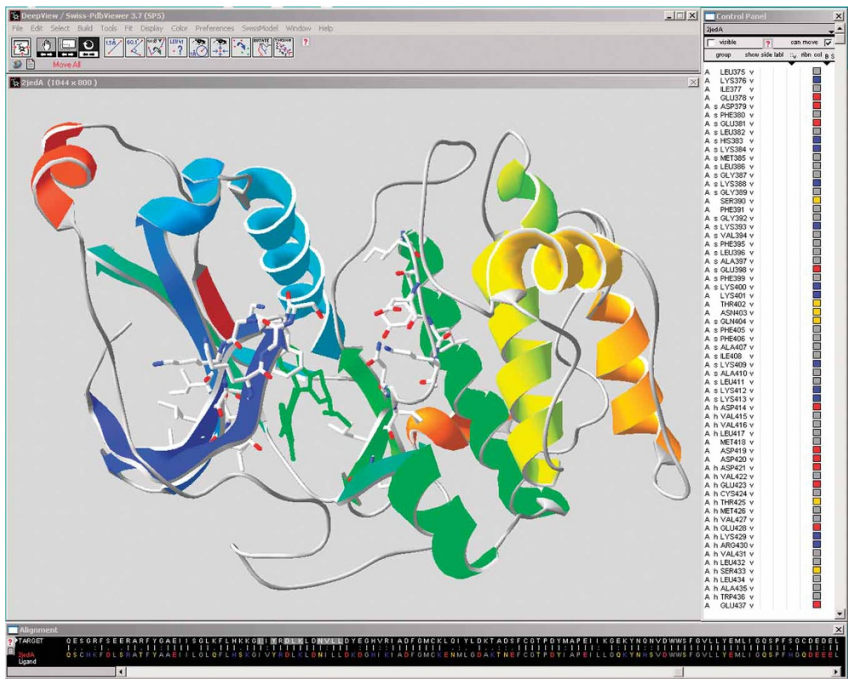

Figure 7 | Model of the kinase domain of the putative PKC δ from *Drosophila* shown as ribbon representation in DeepView colored from blue (N terminus) to red (C terminus). Characteristic residues of the Ser-Thr kinase active site and ATP-binding motif (identified by PROSITE) are shown as sticks. The position of the inhibitor molecule Nvp-Xaa228 in the template structure is highlighted in green.

The two PE/DAG (Phorbol esters/diacylglycerol)-binding domains of the PKC δ protein can also be modeled on the basis of templates identified by iterative profile Blast and HMM-HMM search

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Petitioner Merck, Ex. 1012, p. 11

methods. Few templates were detected by both methods and we decided to use the structure (PDB: 1PTQ[76]) of the second activator-binding domain (PE/DAG) of an orthologous protein (the mouse PKC δ, UniProt AC: P28867) to build the models for the two PE/DAG domains. The alignment between the two PE/DAG domains of the PKC δ and the template is largely unambiguous, and the resulting model is of good quality according to the standard structure assessment tools (Step 8). Particular attention should be paid to the characteristic histidine and cysteine residues, which are assumed to be involved in the coordination of zinc ions. Correctly mapping these residues in the target–template sequence alignment ensures that they have a chemically plausible three-dimensional arrangement in the homology model.

Additional examples can be found in ref. 40 and in the tutorial provided on the SWISS-MODEL workspace website http://swissmodel.expasy.org/workspace/tutorial/.

ACKNOWLEDGMENTS We are grateful to Dr Michael Podvinec for his enthusiastic support and excellent coordination of the Scrum process for the SWISS-MODEL team. We are thankful for financial support of our group by the Swiss Institute of Bioinformatics (SIB).

Published online at http://www.natureprotocols.com/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

1. Berman, H., Henrick, K., Nakamura, H. & Markley, J.L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids. Res.* **35**, D301–D303 (2007).
2. Wu, C.H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids. Res.* **34**, D187–D191 (2006).
3. Chothia, C. Proteins. One thousand families for the molecular biologist. *Nature* **357**, 543–544 (1992).
4. Chothia, C. & Lesk, A.M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).
5. Topham, C.M. *et al.* An assessment of COMPOSER: a rule-based approach to modelling protein structure. *Biochem. Soc. Symp.* **57**, 1–9 (1990).
6. Sali, A. & Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
7. Peitsch, M.C. Protein modelling by e-mail. *BioTechnology* **13**, 658–660 (1995).
8. Tramontano, A. & Morea, V. Assessment of homology-based predictions in CASP5. *Proteins* **53** (Suppl. 6): 352–368 (2003).
9. Tress, M., Ezkurdia, I., Grana, O., Lopez, G. & Valencia, A. Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins* **61** (Suppl. 7): 27–45 (2005).
10. Jauch, R., Yeo, H.C., Kolatkar, P.R. & Clarke, N.D. Assessment of CASP7 structure predictions for template free targets. *Proteins* **69** (Suppl. 8): 57–67 (2007).
11. Kopp, J., Bordoli, L., Battey, J.N., Kiefer, F. & Schwede, T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins* **69** (Suppl. 8): 38–56 (2007).
12. Kryshtafovych, A., Fidelis, K. & Moult, J. Progress from CASP6 to CASP7. *Proteins* **69** (Suppl. 8): 194–207 (2007).
13. Hillisch, A., Pineda, L.F. & Hilgenfeld, R. Utility of homology models in the drug discovery process. *Drug Discov. Today* **9**, 659–669 (2004).
14. Kopp, J. & Schwede, T. Automated protein structure homology modeling: a progress report. *Pharmacogenomics* **5**, 405–416 (2004).
15. Marti-Renom, M.A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
16. Peitsch, M.C. About the use of protein models. *Bioinformatics* **18**, 934–938 (2002).
17. Tramontano, A. In *Computational Structural Biology* (eds. Schwede T. & Peitsch M.C.) (World Scientific Publishing, Singapore, 2008).
18. Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
19. Soto, C.S., Fasnacht, M., Zhu, J., Forrest, L. & Honig, B. Loop modeling: sampling, filtering, and scoring. *Proteins* **70**, 834–843 (2008).
20. Rohl, C.A., Strauss, C.E., Chivian, D. & Baker, D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**, 656–677 (2004).
21. Fiser, A., Do, R.K. & Sali, A. Modeling of loops in protein structures. *Protein Sci.* **9**, 1753–1773 (2000).
22. Canutescu, A.A., Shelenkov, A.A. & Dunbrack, R.L. Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001–2014 (2003).
23. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94 (1999).
24. Dunbrack, R.L. Jr. Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.* **16**, 374–384 (2006).
25. Sommer, I., Toppo, S., Sander, O., Lengauer, T. & Tosatto, S.C. Improving the quality of protein structure models by selecting from alignment alternatives. *BMC Bioinformatics* **7**, 364 (2006).
26. Tress, M.L., Jones, D. & Valencia, A. Predicting reliable regions in protein alignments from sequence profiles. *J. Mol. Biol.* **330**, 705–718 (2003).
27. Vingron, M. Near-optimal sequence alignment. *Curr. Opin. Struct. Biol.* **6**, 346–352 (1996).
28. Melo, F. & Feytmans, E. Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.* **277**, 1141–1152 (1998).
29. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883 (1990).
30. Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**, 2714–2726 (2002).
31. Fasnacht, M., Zhu, J. & Honig, B. Local quality assessment in homology models using statistical potentials and support vector machines. *Protein Sci.* **16**, 1557–1568 (2007).
32. Wallner, B. & Elofsson, A. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci.* **15**, 900–913 (2006).
33. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291 (1993).
34. Hooft, R.W., Vriend, G., Sander, C. & Abola, E.E. Errors in protein structures. *Nature* **381**, 272 (1996).
35. Aloy, P., Pichaud, M. & Russell, R.B. Protein complexes: structure prediction challenges for the 21st century. *Curr. Opin. Struct. Biol.* **15**, 15–22 (2005).
36. Alber, F. *et al.* Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).
37. Junne, T., Schwede, T., Goder, V. & Spiess, M. The plug domain of yeast Sec61p is important for efficient protein translocation, but is not essential for cell viability. *Mol. Biol. Cell* **17**, 4063–4068 (2006).
38. Battey, J.N. *et al.* Automated server predictions in CASP7. *Proteins* **69** (Suppl. 8): 68–82 (2007).
39. Koh, I.Y. *et al.* EVA: evaluation of protein structure prediction servers. *Nucleic Acids. Res.* **31**, 3311–3315 (2003).
40. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195–201 (2006).
41. Eswar, N. *et al.* Tools for comparative protein structure modeling and analysis. *Nucleic Acids. Res.* **31**, 3375–3380 (2003).
42. Bates, P.A., Kelley, L.A., MacCallum, R.M. & Sternberg, M.J. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* (Suppl. 5): 39–46 (2001).
43. Fernandez-Fuentes, N., Madrid-Aliste, C.J., Rai, B.K., Fajardo, J.E. & Fiser, A. M4T: a comparative protein structure modeling server. *Nucleic Acids Res.* **35**, W363–W368 (2007).
44. Fox, J.A., McMillan, S. & Ouellette, B.F. Conducting research on the web: 2007 update for the bioinformatics links directory. *Nucleic Acids Res.* **35**, W3–W5 (2007).
45. Schwede, T., Diemand, A., Guex, N. & Peitsch, M.C. Protein structure computing in the genomic era. *Res. Microbiol.* **151**, 107–112 (2000).
46. Kopp, J. & Schwede, T. The SWISS-MODEL repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.* **32**, D230–D234 (2004).
47. Schwede, T., Kopp, J., Guex, N. & Peitsch, M.C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31**, 3381–3385 (2003).
48. Guex, N. & Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).

Petitioner Merck, Ex. 1012, p. 12

49. Andreeva, A. *et al.* SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32**, D226–D229 (2004).
50. Greene, L.H. *et al.* The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.* **35**, D291–D297 (2007).
51. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
52. Zdobnov, E.M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
53. Mulder, N.J. *et al.* New developments in the InterPro database. *Nucleic Acids Res.* **35**, D224–228 (2007).
54. Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
55. Jones, D.T. & Ward, J.J. Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **53** (Suppl. 6): 573–578 (2003).
56. Jones, D.T., Taylor, W.R. & Thornton, J.M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**, 3038–3049 (1994).
57. Fink, A.L. Natively unfolded proteins. *Curr. Opin. Struct. Biol.* **15**, 35–41 (2005).
58. Radivojac, P. *et al.* Intrinsic disorder and functional proteomics. *Biophys. J.* **92**, 1439–1456 (2007).
59. Dyson, H.J. & Wright, P.E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
60. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
61. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **33** Database Issue: D39–D45 (2005).
62. Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
63. Muller, C.W., Schlauderer, G.J., Reinstein, J. & Schulz, G.E. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **4**, 147–156 (1996).
64. Söding, J., Biegert, A. & Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–248 (2005).
65. Benkert, P., Tosatto, S.C. & Schomburg, D. QMEAN: a comprehensive scoring function for model quality assessment. *Proteins* **71**, 261–277 (2008).
66. van Gunsteren, W.F. *et al. Biomolecular Simulations: the GROMOS96 Manual and User Guide* (VdF Hochschulverlag ETHZ, Zürich, 1996).
67. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
68. Hulo, N. *et al.* The PROSITE database. *Nucleic Acids Res.* **34**, D227–D230 (2006).
69. Stivers, J.T. & Jiang, Y.L. A mechanistic perspective on the chemistry of DNA repair glycosylases. *Chem. Rev.* **103**, 2729–2759 (2003).
70. Seeberg, E., Eide, L. & Bjoras, M. The base excision repair pathway. *Trends Biochem. Sci.* **20**, 391–397 (1995).
71. Alseth, I. *et al.* A new protein superfamily includes two novel 3-methyladenine DNA glycosylases from Bacillus cereus, AlkC and AlkD. *Mol. Microbiol.* **59**, 1602–1609 (2006).
72. Groves, M.R. & Barford, D. Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9**, 383–389 (1999).
73. Dalhus, B. *et al.* Structural insight into repair of alkylated DNA by a new superfamily of DNA glycosylases comprising HEAT-like repeats. *Nucleic Acids Res.* **35**, 2451–2459 (2007).
74. Nishizuka, Y. Membrane phospholipid degradation and protein kinase C for cell signalling. *Neurosci. Res.* **15**, 3–5 (1992).
75. Mellor, H. & Parker, P.J. The extended protein kinase C superfamily. *Biochem. J.* **332** (Part 2): 281–292 (1998).
76. Zhang, G., Kazanietz, M.G., Blumberg, P.M. & Hurley, J.H. Crystal structure of the cys2 activator-binding domain of protein kinase C delta in complex with phorbol ester. *Cell* **81**, 917–924 (1995).
77. Henrick, K. & Thornton, J.M. PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361 (1998).
78. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).

© 2009 Nature Publishing Group http://www.nature.com/natureprotocols

Petitioner Merck, Ex. 1012, p. 13

Corrigendum: Protein structure homology modeling using SWISS-MODEL workspace

Lorenza Bordoli, Florian Kiefer, Konstantin Arnold, Pascal Benkert, James Battey & Torsten Schwede

Nat. Protoc. **4**, 1–13 (2009); doi:1038/nprot.2008.197; published online 11 December 2008; corrected online 18 June 2009.

The version of this article initially published indicated that only Torsten Schwede was affiliated with the Swiss Institute of Bioinformatics in addition to the Biozentrum, University of Basel, Basel, Switzerland. However, all six authors are affiliated with both the Biozentrum and the Swiss Institute of Bioinformatics. The error has been corrected in the HTML and PDF versions of the article.

Petitioner Merck, Ex. 1012, p. 14