The Wayback Machine - https://web.archive.org/web/20110519141121/http://swissmodel.expasy.org/?pid=smh01&uid=&token=

Swiss Institute of Bioinformatics | BIOZENTRUM Universität Basel

SWISS-MODEL

Modelling

myWorkspace

Automated Mode

Alignment Mode

Project Mode

Tools

Template Identification

Domain Annotation

Structure Assessment

Template Library

Repository

Search by Sequence

Search by AC

Search by full text

Documentation

SWISS-MODEL Workspace

SWISS-MODEL Repository

Structures & Models

Helpdesk

SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists worldwide.

What's new?

- New **Beta SWISS-MODEL** pipeline for automated model building with oligomers and ligands is now available for testing.
- Find more news on **SWISS-MODEL Blog**
- ... faster news on **Twitter**
- Follow us on **Facebook**

SWISS-MODEL Team

Torsten Schwede:	Project Leader
Florian Kiefer:	SWISS-MODEL Repository
Lorenza Bordoli:	Method Development and user support
Konstantin Arnold:	SWISS-MODEL Workspace

References:

When you publish or report results using SWISS-MODEL, please cite the relevant publications:

- Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 22,195-201.
- Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Research. 37, D387-D392.
- Peitsch, M. C. (1995) Protein modeling by E-mail Bio/Technology 13: 658-660.

Petitioner Merck, Ex. 1066, p. 1

SWISS-MODEL is developed by the Protein Structure Bioinformatics group at the SIB - Swiss Institute of Bioinformatics & the Biozentrum University of Basel. © 2011.

Petitioner Merck, Ex. 1066, p. 2

The Wayback Machine - https://web.archive.org/web/20110519140410/http://swissmodel.expasy.org/?pid=smd11&uid=&token=

SWISS-MODEL

Introduction to SWISS-MODEL Workspace

The SWISS-MODEL Workspace is a web-based integrated service dedicated to protein structure homology modelling. It assists and guides the user in building protein homology models at different levels of complexity.

Building a homology model comprises four main steps: identification of structural template(s), alignment of target sequence and template structure(s), model building, and model quality evaluation. These steps can be repeated until a satisfying modelling result is achieved. Each of the four steps requires specialized software and access to up-to-date protein sequence and structure databases.

- **Help** More detailed help section.
- **Tutorial** explaining how to use SWISS-MODEL Workspace
- **Protocol** how to use Protein structure homology modeling using SWISS-MODEL workspace on Nature Protocols
- **References** SWISS-MODEL Workspace Reference List

Petitioner Merck, Ex. 1066, p. 3

SWISS-MODEL is developed by the Protein Structure Bioinformatics group at the SIB - Swiss Institute of Bioinformatics & the Biozentrum University of Basel. © 2011.

Petitioner Merck, Ex. 1066, p. 4

The Wayback Machine - https://web.archive.org/web/20110519144707/http://swissmodel.expasy.org/?pid=smd02&uid=&token=

SWISS-MODEL

Introduction to SWISS-MODEL Workspace

The SWISS-MODEL Workspace is a web-based integrated service dedicated to protein structure homology modelling. It assists and guides the user in building protein homology models at different levels of complexity.

Building a homology model comprises four main steps: identification of structural template(s), alignment of target sequence and template structure(s), model building, and model quality evaluation. These steps can be repeated until a satisfying modelling result is achieved. Each of the four steps requires specialized software and access to up-to-date protein sequence and structure databases.

Protein sequence and structure databases necessary for modelling are accessible from the workspace and are updated in regular intervals. Software tools for *template selection*, *model building*, and structure *quality evaluation* can be invoked from within the workspace.

A personal working environment (*workspace*), where several modelling projects can be carried out in parallel, is provided for each user.

This help file provides references and illustrate the use of the individuals tools available from within the SWISS-MODEL Workspace.
A tutorial to facilitate the first steps of working with SWISS-MODEL Workspace as a list of most frequently asked questions is provided here: Tutorial
Please also take a look at the following published [Protocol]

Workspace

The SWISS-MODEL Workspace provides a personal web-based area for each user in which protein homology models can be built and the results of completed modelling projects are stored and visualized.

In the workspace a list of the current modeling work units and their current status is displayed: *submitted* (the job has been submitted to the pipeline but still queuing), *running* (job is running and programs are calculating), *finished* (job has been completed, final results are available) or *failed/stopped* (if something went wrong during the process).

Depending on the type of job the user has submitted a different tag will be associated with a work unit: *Template Identification* for template identification, *Sequence Scanning* for secondary structure and disorder prediction and domain assignment, *Structure Assessment* for structure quality assessement. And *Modelling Automatic*, *Modelling Project*, *Modelling Alignment* respectively for automated, alignemnt or project mode modeling requests.

After completion of the modelling procedure (~ a few minutes up to several hours), the results are stored in the workspace and the user is notified about the completion.The user can access the results output by clicking on the work unit ID number.

The results are stored for one week on the server. The remainig time before deletion of a given work unit is also displayed. The user can decide to either delete a work unit or to prolonge its life span by clicking on the corresponding link.

Beware: Each user can submit up to a maximum of 25 work units.

Domain assignment, Secondary Structure and Disorder Prediction

Many proteins are modular and made up of several structurally distinct domains, which often reflect evolutionary relationships and may correspond to units of molecular function.The sensitivity and performance of profile-based template search methods can often be improved when the template search is performed on individual domains rather than the whole target sequence. IprScan (see below) allows for protein domains and functional site prediction.
Protein disorder prediction measures and displays the propensity of protein sequences to be ordered or disordered. The result can aid the assignment of templates to a specific region of the target protein by complementing the IprScan approach to globular domains and feature discovery.
Secondary structure prediction methods are especially useful when combined with other types of analyses: e.g. in cases where only templates with very low sequence homology can be detected by sequence-based search methods, predicted secondary structure may help to decide if a putative template shares structural features of the target protein.

InterPro Domain Scan

The member databases of InterPro (*Mulder et al.*) allow for both the identification of protein domains and the assignment of protein function. Using the InterPro Domain Scan (IprScan, *Zdobnov et al.*), protein domains and functional sites can be assigned to regions of a target sequence.

The following databases are currently part of the InterPro Domain scan method:

HMMPfam: Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families.

HMMTigr: TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology.

Petitioner Merck, Ex. 1066, p. 5

ProfileScan: PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs. There are a number of protein families as well as functional or structural domains that cannot be detected using patterns (see below) due to their extreme sequence divergence. The use of techniques based on weight matrices (also known as profiles) allows the detection of such domains.

SuperFamily: SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure, based on SCOP.

BlastProDom: The ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches. The ProDom database has been designed as a tool to help analyze domain arrangements of proteins and protein families.

FPrintScan: PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family.

HMMSmart:SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures.

ScanRegExp:PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs. Some biologically significant amino acid patterns can be summarised in the form of regular expressions.

The results of the InterPro domain mapping is displayed in combination with the alignment to putative template structures, allowing the user to identify template structures spanning one or more domains of the target protein. For low homology templates, the IprScan functional site annotation of the target sequence can be used to verify that putative templates share essential functional features. The InterPro functional annotations for individual template structures are accessible from the workspace as links to the SMTL library and external resources..

PsiPred Secondary Structure Prediction

PSIPRED is a method for protein secondary structure prediction (*Jones DT et al.*).

The plot shows position in the sequence against probability of being be part of a alpha helix (H) , extended beta strand (E) or a coil region (C). The result of the prediction is plottet on the x-axis of the plot.

DISOPRED Disorder Prediction

DISOPRED (v 2) is a neural-network based predictor of disordered regions in proteins (*Jones DT et al.*).

The majority of water-soluble proteins have structures that are globular and relatively static. However, some proteins have regions that are natively disordered. Disordered regions are flexible, dynamic and can be partially or completely extended in solution. Native disorder also exists in global structures such as extended random coil proteins with negligible secondary structure or molten globules, which have regular secondary structure elements but have not condensed into a stable globular fold.The primary function of disorder appears to be molecular recognition of proteins and nucleic acids. It has been speculated that the multiple metastable conformations, adopted by disordered binding sites, allows recognition of several targets with high specificity and low affinity. Order to disorder transitions also provide a mechanism for controlling protein concentration via proteolytic degradation.

The plot shows position in the sequence against probability of being disordered (from 0 to 1). The 'filter' curve represents the outputs from DISOPRED and the 'output' curve the outputs from a linear SVM classifier (DISOPREDsvm). The outputs from DISOPREDsvm are included to indicate shorter, low confidence predictions of disorder.

Asterisks (*) represent disordered predictions and dots (.) prediction of order.

The disopred predictions are given at a default false positive rate threshold of 2%. But this value can be changed by the user.

MEMSAT

MEMSAT 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.

Template Identification

The degree of difficulty in identifying a suitable template for a target sequence can range from "trivial" for well-characterized protein families to "impossible" for proteins with an unknown fold. The SWISS-MODEL Workspace provides access to a set of increasingly complex and computationally demanding methods to search for templates.

Templates which are close homologues of the target can be identified using a gapped BLAST (*Altschul et al.*) query against the ExPDB template library extracted from PDB.

Options for the **BLAST** database search are:
E-value cutoff: sets the threshold expectation value for keeping alignments. It describes how often a given score is expected to occur random;
Matrix: the protein substitution matrix;
SEG Filter: filters the query sequence for low-complexity subsequences;
Descriptions: sets the number of database sequences for which to show the one-line summary descriptions at the top of a BLAST report;
Alignments: truncates the report to the selected number of alignments;

Petitioner Merck, Ex. 1066, p. 6

When no suitable templates are identified, or only parts of the target sequence are covered, two additional approaches for the sensitive detection of distant relationships among protein families are provided:

Iterative Profile Blast: the template library is searched with PSI-BLAST (*Altschul et al.*) using an iteratively generated sequence profile based on NR (*Wheeler et al.*). This method has been initially introduced as PDB-Blast by Godzik and coworkers.

- The first run searches the NR database and derive a profile for the query sequence. The following options are available:
Iterations: number of iteration for the NR database search and profile (PSSM) generation;
Matrix: the protein substitution matrix;
Evalue: The E-value threshold for inclusion in PSSM. All alignments better than this threshold are used in constructing the PSSM;
SEG Filter: filters the query sequence for low-complexity subsequences;

- Then with this profile, the final run searches the SWISS-MODEL template library (ExPDB). The following options are available:
Database to search: Clustered versions of ExPDB (e.g. ExPDB90, sequences clustered to 90% of redundancy) which combine closely related sequences into a single record;
E-value cutoff: sets the threshold expectation value for keeping alignments. It describes how often a given score is expected to occur random;
Matrix: the protein substitution matrix;
SEG Filter: filters the query sequence for low-complexity subsequences;
Descriptions: sets the number of database sequences for which to show the one-line summary descriptions at the top of a BLAST report;
Alignments: truncates the report to the selected number of alignments;

HHSearch: To detect distantly related template structures, a target sequence can be searched against a Hidden Markov Model (HMM) based template library. Each HMM of the library is based on a multiple sequence alignment of the template sequence built by PSI-BLAST search (against nr90 & nr70) enriched with secondary structure assignment.
In analogy a HMM is built for the target sequence, which is subsequently used to search against the template library. Only alignments which score more than a given P-value cut-off are reported. Model building and library searches are performed using the HHSEARCH (v. 1.5.01) software package (*Söding et al.*) with default parameters.
For detailed documentation, please visit the official HHSEARCH site [http://toolkit.tuebingen.mpg.de/hhpred]

Display of template identification results

A condensed graphical view of the modeling task is provided containing the target sequence, the template matches sorted and colored according to the associated E-value, and the InterPro mappings. Clickable bars indicate the matched regions and guide the user to the underlying original program output.

In the InterPro output a link leads to the detailed InterPro page for this entry.

In the output of the different template identification programs the template annotations (via the link to the SWISS-MODEL Template library) and target-template alignment can be retrieved.

Alignments can be obtained as DeepView project file. The latter allows the user to visualize the different alignments in the structural context of the template, to correct misplaced insertions and deletions, and to manually adjust misaligned regions. The modified project can then be saved to disk and submitted as "project mode" to the workspace for model building by the SWISS-MODEL pipeline.

When searching a clustered version of the SWISS-MODEL Template library (e.g. ExPDB90) only the alignment between the target sequence and the sequence of the representative of the cluster is shown. Information about the members of the cluster is presented in the detailed output of the different template search programs. For each template, the SWISS-MODEL workspace provides a summary showing a small ribbon representation, experimental details, information about bound molecules, as well as links to PDB (*Westbrook et al.*), SCOP (*Andreeva et al.*), CATH (*Pearl et al.*), PDBsum (*Laskowskiet et al.*), and MSD (*Velankaret et al.*).

Model building

Depending on the difficulty of the modelling task, three different types of modelling requests (*automated mode*, *alignment mode*, *project mode*) are provided, which differ in the amount of user intervention.

Modelling requests are computed by the SWISS-MODEL server homology modelling pipeline (*Schwede et al.*).

Automated Mode

The "automated mode" is suited for cases where the target-template similarity is sufficiently high to allow for fully automated modelling. As a rule of thumb, automated sequence alignments are sufficiently reliable when target and template share more than 50% percent of sequence identity.

This submission requires only the amino acid sequence or the UniProt accession code of the target protein as input data. The pipeline will automatically select suitable templates based on a Blast (*Altschul et al.*) E-value limit (which can be adjusted upon submission), experimental quality, bound substrate molecules, or different conformational states of the template.

Depending on the planned model application, it can be necessary to select a different structural template than the one ranked first in the automated process. Typical examples are proteins in different conformational states, e.g. 1ake vs. 4ake. It is possible to specify the structure to be used as modelling template either by identifying an entry in the SWISS MODEL Template library by PDB-ID + ChainID e.g. "1ake" chain "A", or by uploading a file in PDB format (*) with coordinates of the template structure. Please make sure that this file contains only a single protein chain, and does not contain chemically modified amino acids, hereto atoms, ligands, etc.

(*) A simple PDB-like file containing the coordinates of the template structure. For more information about PDB file format please see link: http://www.wwpdb.org/docs.html

Alignment Mode

Petitioner Merck, Ex. 1066, p. 7

Multiple sequence alignments are a common tool in many molecular biology projects. If the three-dimensional structure is known for at least one of the members, this alignment can be used as starting point for comparative modelling using the "alignment mode".
The "alignment mode" allows the user to test several alternative alignments and evaluate the quality of the resulting models in order to achieve an optimal result.

In order to facilitate the use of alignments in different formats, the submission is implemented as a three step procedure:

1. Prepare a multiple sequence alignment.

- It must contain at least your target sequence and the template sequence
- Use any of your favorite alignment tools. We recommend T_COFFEE by Cedric Notredame
- Make sure the sequence names are "reasonable"

2. Submit your alignment to the Workspace Alignment Mode.

- Possible formats are: FASTA, MSF, CLUSTALW, PFAM and SELEX
- You may either upload your file or cut & paste
- Don't forget to specify the correct alignment format
- Here is a small example for testing (cut & paste):

```
CLUSTAL W (1.82) multiple sequence alignment

THN_DENCL       KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH- 46
THNX_TEST       KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS-YPNK 46
1crnA           TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN- 46
                .:***  ..*  :  **: * .. :**  :** **..: **  *
```

3. Select Target and Template

- The alignment (as it was interpreted by the server) should now be displayed in the bottom part of the page.
- The script will try to make a good guess for the correct names based on your submission.
- Select the sequence name of the target sequence (e.g. THN_DENCL)
- Select the sequence of the template structure (e.g. 1crnA). You don't need to use PDB IDs, you may use any name you like.
- Specify the template structure to which this sequence belongs. This template MUST be part of the ExPDB template library. Please use the SWISS-MODEL Template library tool to check...
- Don't forget to specify the correct CHAIN ID. Note that PDB's chain IDs are normally in capital letters.

```
Target   sequence: THN_DENCL
Template sequence: 1crnA      PDB-Code: 1crn      Chain-ID:  A
```

4. Check Alignment and Submit

- The alignment at the bottom of the page should represent the correct mapping of the template structure on the target sequence. Please check carefully before submission.
- As usual, please provide name and e-mail for the SWISS-MODEL submission.
- Good Luck with your model ....

The server pipeline will build the model purely based on this alignment. During the modelling process, implemented as rigid fragment assembly in the SWISS-MODEL (*Schwede et al.*) pipeline, the modelling engine might introduce minor heuristic modifications to the placement of insertions and deletions.

Supported Alignment formats

The following formats are currently supported: FASTA, MSF, CLUSTALW, PFAM and SELEX;

Examples:

fasta:

```
>THN_DENCL
KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH-
>THNX_TEST
KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS-YPNK
>1crnA
TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN-
```

clustal:

```
CLUSTAL W (1.82) multiple sequence alignment

THN_DENCL       KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH- 46
THNX_TEST       KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS-YPNK 46
1crnA           TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN- 46
                .:***  ..*  :  **: * .. :**  :** **..: **  *
```


Petitioner Merck, Ex. 1066, p. 8

msf:

```
 !!AA_MULTIPLE_ALIGNMENT 1.0

  thn_dencl.msf MSF:  47 Type: P 08/08/05 CompCheck:  427 ..

  Name: THN_DENCL  Len: 47  Check: 8212 Weight: 1.00
  Name: THNX_TEST  Len: 47  Check: 5295 Weight: 1.00
  Name: 1crnA      Len: 47  Check: 6920 Weight: 1.00

//

           1                                           47
THN_DENCL  KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH~
THNX_TEST  KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS.YPNK
1crnA      TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN~
```

Project Mode

In difficult modeling situations, where the correct alignment between target and template cannot be clearly determined by sequence based methods, visual inspection and manual manipulation of the alignment can significantly help improving the quality of the resulting model.

Project files contain the superposed template structures, and the alignment between the target and template. Project files can be generated inside the program DeepView (Swiss-PdbViewer *Guex et al.*), by the workspace template selection tools, and are also the default output format of the modeling pipeline. This allows analyzing and iteratively improving the the models generated by the "Automated mode" and "Alignment mode" modeling approaches.

The program DeepView can be downloaded freely from the tools section or from the ExPASy web site .

DeepView

The program DeepView (Swiss-PdbViewer, *Guex et al.*) can be used to generate, display, analyze and manipulate modeling project files for the SWISS-MODEL workspace.

Project files contain the superposed template structures, and the alignment between the target and template. The user has therfor full control over essential modelling parameters, i.e. the choice of template structures, the correct alignment of residues, and the placement of insertions and deletions in the context of the three-dimensional structure.

Project files can be generated inside DeepView, by the workspace template identification tools, and are also the default output format of the modeling pipeline. This allows analyzing and iteratively improving the output of the different modeling tools.

DeepView allows to visualize the model and the templates, and to analyse certain structural features e.g. Ramachandran plots or electrostatic properties. Moreover, it allows adjusting manually the placement of insertions and deletions in the alignment on which the initial modelling process was based on. The project with the modified alignment can then be re-submitted to the SWISS-MODEL workspace for model building.

DeepView can be downloaded at: http://www.expasy.org/spdbv/

DeepView does not require administrator privileges for installation. E.g. under MS windows, simply uncompress the distributed archive at any location you like (e.g. c:\spdbv or on your desktop) and start working by starting the spdbv.exe application.

Input target sequence and UniProt AC code

The amino acid sequence of a protein to be modeled or analyzed can be submitted in FASTA or raw format. If the protein sequence is deposited in the UniProt (*Bairoch et al.*)knowledgebase, the AC (ACcession number) for the entry can be also specified.

Examples:

- raw format: the amino acids sequence of the protein in plain-text:

```
MVEIVYWSGTGNTEAMANEIEAAVKAAGADVESVRFEDTNVDDVASKDVILLGCPAMGSE
ELEDSVVEPFFTDLAPKLKGKKVGLFGSYGWGSGEWMDAWKQRTEDTGATVIGTAIVNEM
PDNAPECKELGEAAAKA
```

- FASTA format consists of a single-line description, followed by lines of sequence data. The first character of the description line is a greater-than (">") symbol:

Petitioner Merck, Ex. 1066, p. 9

```
>sp|P00321|FLAV_MEGEL Flavodoxin - Megasphaera elsdenii.
MVEIVYWSGTGNTEAMANEIEAAVKAAGADVESVRFEDTNVDDVASKDVILLGCPAMGSE
ELEDSVVEPFFTDLAPKLKGKKVGLFGSYGWGSGEWMDAWKQRTEDTGATVIGTAIVNEM
PDNAPECKELGEAAAKA
```

- UniProt Accesion number: P00321

Display of modeling results

Coordinates of the model, the underlying alignment, log files, and quality evaluations can be accessed and downloaded via web-browser from the workspace.

Model Details

This section gives access to display the model and download its coordinates.

The model coordinates are available in two different formats:

- DeepView project files (recommended).
- PDB format

PDB formatted protein models can be displayed by any molecular visualization tool or browser-plugin. Here is a short list of freely available software:

- DeepView (MS Windows, Macintosh, Linux)
- DINO (Linux, IRIX, OSF,SUN)
- Rasmol (MS Windows, Mac, Unix)
- CHIME Plugin (requires registration)

If the model has been build using the *Automated Mode*, information about the template(s) used for modeling is provided with cross references to structural information databases via the link to the SWISS MODEL Template library.

Alignment Output

Displays the target template sequence alignment used in the modeling procedure and the assigned secondary structure.

Modeling Log

The modeling log gives a detailed description of the individual modeling steps. The models are built using the SWISS-MODEL server pipeline (*Schwede et al.*). The modelling log shows the individual steps during model building *Guex et al.*, especially which parts of the model have been built ab initio (i.e. insertions / deletions).

Template SelectionLog

The logfile provides information about the template selection step to search the SWISS-MODEL Template library for suitable templates.

Protein Structure & Model Assessment Tools

Evaluation of model quality is a crucial step in homology modeling. While the performance of the automated SWISS-MODEL (*Schwede et al.*) pipeline in general is continuously evaluated by the EVA project (*Koh et al.*), the quality of individual models can vary significantly.

Therefore, graphical plots of Anolea mean force potential (*Melo et al.*) and GROMOS empirical force field energy (*van Gunsteren et al.*) are provided to enable the user to estimate the local quality of the predicted structure. The stereochemistry of protein models and template structures can be analysed with Whatcheck (*Hooft et al.*) and Procheck (*Laskowski et al.*). In order to be able to rank alternative models of the same target, pseudo energies for the entire model as calculated by QMEAN (*Benkert et al.*) and DFIRE (*Zhou et al.*) are provided as well. To facilitate the description of template and model structures, DSSP (*Kabsch et al.*) and Promotif (*Hutchinson et al.*) can be invoked to classify structural features.

Anolea

The atomic empirical mean force potential ANOLEA (*Melo et al.*) is used to assess packing quality of the models. The program performs energy calculations on a protein chain, evaluating the "Non- Local Environment" (NLE) of each heavy atom in the molecule.

The y-axis of the plot represents the energy for each amino acid of the protein chain. Negative energy values (in green) represent favourable energy environment whereas positive values (in red) unfavourable energy environment for a given amino acid.

QMEAN

QMEAN (*Benkert et al.*) is a composite scoring function for the estimation of the global model quality. QMEAN consisting of five structural descriptors: The local geometry is analysed by a torsion angle potential over three consecutive amino acids. A distance-dependent pairwise residue-level potential is used to assess long-range interactions. A solvation potential describes the burial status of the residues. Two simple terms describing the agreement of predicted and calculated secondary structure and solvent accessibility, respectively, are also included.

QMEAN returns a pseudo energy of the whole model which can be used in order to compare and rank alternative models of the same target. The

Petitioner Merck, Ex. 1066, p. 10

lower the predicted energy, the better model. Additionally, the pseudo energies of the five contributing terms are provided.

DFire

DFIRE (*Zhou et al.*) is an all-atom statistical potential based on a distance-scaled finite ideal-gas reference state. DFIRE is used to assess non-bonded atomic interactions in the protein model.

A pseudo energy for the entire model is provided which reflects the quality of the model and can be used for ranking alternative predictions of the same target. A lower energy indicates that a model is closer to the native conformation.

Gromos

GROMOS (*van Gunsteren et al.*) is a general-purpose molecular dynamics computer simulation package for the study of biomolecular systems and can be applied to the analysis of conformations obtained by experiment or by computer simulation.

The y-axis of the plot represents the energy for each amino acid of the protein chain. Negative energy values (in green) represent favourable energy environment whereas positive values (in red) unfavourable energy environment for a given amino acid.

What Check

What Check comprises several tools for protein structure verification (*Hooft et al.*).

Procheck

The PROCHECK suite of programs (*Laskowski et al.*) assess the "stereochemical quality" of a given protein structure. The aim of PROCHECK is to assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures.

PROMOTIF

PROMOTIF (*Hutchinson et al.*) automatically identifies, classify and analyse a number of supersecondary structural motifs in proteins. Any resulting patterns will be useful in prediction of protein structure from amino acid sequence. Motifs analyzed include beta turns, gamma turns, Greek keys, beta hairpins and beta bulges. Data from PROMOTIF analyses are included in the PDBsum (*Laskowskiet et al.*) web site, which provides information derived from all currently available protein coordinate files.

DSSP

The DSSP (*Kabsch et al.*) program defines secondary structure, geometrical features and solvent exposure of proteins, given atomic coordinates in Protein Data Bank format. The program does NOT PREDICT protein structure.
The DSSP code

H = alpha helix
B = residue in isolated beta-bridge
E = extended strand, participates in beta ladder
G = 3-helix (3/10 helix)
I = 5 helix (pi helix)
T = hydrogen bonded turn
S = bend

SwissModel Template Library (ExPDB)

The template structure database used by SWISS-MODEL (SMTL or ExPDB library) is derived from the Protein Data Bank (*Westbrook et al.*). In order to allow sequence-based template searches, each PDB entry is split into individual chains. The separated template chains are annotated with information about experimental method, resolution (if applicable), ANOLEA mean force potential (*Melo et al.*), Gromos96 energy (*van Gunsteren et al.*) and PQS (*Henrick et al.*) quaternary state assignment to allow for rapid retrieval of the relevant structural information during template selection. Theoretical models, structures only consisting of C alpha atoms and irregularly formatted database entries are removed.
In order to speed up the sequence database search step of the template identification algorithms and to provide a clear and concise overview of the results, templates sharing 100% sequence identity are grouped into a SMTL100 library using the program CD-HIT, a fast clustering method for sequences at high identity thresholds (*Li et al.*). Clusters of sequences having 90%, 70% and 50% sequence identity are derived from the RCSB non-redundant PDB lists.

The ExPDB codes are constructed according to the following rule: PDBCODE+ChainID

Examples:
- *Light harvesting protein*: 1cpc contains two chains (with IDs A & B).
The corresponding ExPDB entries are respectively:

- Chain A: 1cpcA
- Chain B: 1cpcB

User specified template:

Depending on the planned model application, it can be necessary to select a different structural template than the one ranked first in the automated process. Typical examples are proteins in different conformational states, e.g. 1ake vs. 4ake. It is possible to specify the structure to be used as

Petitioner Merck, Ex. 1066, p. 11

modelling template either by identifying an entry in the SWISS-MODEL template library by PDB-ID + ChainID e.g. "1ake" chain "A", or by uploading a file in PDB format (*) with coordinates of the template structure. Please make sure that this file contains only a single protein chain, and does not contain chemically modified amino acids, hereto atoms, ligands, etc.

(*) A simple PDB-like file containing the coordinates of the template structure. For more information about PDB file format please see link: http://www.wwpdb.org/docs.html

SWISS-MODEL is developed by the Protein Structure Bioinformatics group at the SIB - Swiss Institute of Bioinformatics & the Biozentrum University of Basel. © 2011.

Petitioner Merck, Ex. 1066, p. 12

The Wayback Machine - https://web.archive.org/web/20110519144916/http://swissmodel.expasy.org/?pid=smd03&uid=&token=

SWISS-MODEL

Introduction

The SWISS-MODEL Workspace is a web-based integrated service dedicated to protein structure homology modelling. It assists and guides the user in building protein homology models at different levels of complexity.

Successful model building requires at least one experimentally determined 3D structure (template) that shows significant amino acid sequence similarity with the target sequence. Building a homology model comprises four main steps: identification of structural template(s), alignment of target sequence and template structure(s), model building, and model quality evaluation. These steps can be repeated until a satisfying modelling result is achieved. Each of the four steps requires specialized software and access to up-to-date protein sequence and structure databases.

Protein sequence and structure databases necessary for modelling are accessible from the workspace and are updated in regular intervals. Software tools for *template selection*, *model building*, and structure *quality evaluation* can be invoked from within the workspace. A personal working environment (*workspace*), where several modelling projects can be carried out in parallel, is provided for each user.

The following tutorial aims to facilitate the first steps of working with SWISS-MODEL Workspace. Please let us know if you would like to see other features explained in this tutorial (help-swissmodel@unibas.ch).

How do I work with SWISS-MODEL workspace ?

- How can I create an account?
- How can I manage my projects?

How can I build an homology model using SWISS-MODEL workspace ?

- My protein is quite large, and would like to identify individual domains I could model separately.
- If I have no idea about possible templates for my target, and
  - I want to identify possible templates.
  - I want to try the fully the automatic mode.
- I already have a target-template alignment and want to model my protein
  - by using the alignment submission tool.
  - by visually verifying the alignment in DeepView.

How can I assess protein structure model quality ?

- What accuracy can I expect for a model build by the automated mode of SWISS-MODEL?
- How can I assess protein structure model quality with empirical force-field and Mean Force Potential methods?
- How can I assess geomoetrical accuracy of a structure or model?

---

1. How do I work with SWISS-MODEL workspace ? => How can I create an account?

The SWISS-MODEL Workspace provides a personal web-based area for each user in which protein homology models can be built and the results of completed modelling projects are stored and visualized. It is not necessary to create an account; you may continue to use SWISS-MODEL as before by just providing an email address in the submission form, or by bookmarking the submission window. However, you will not be able to manage your projects inside the Workspace, and we therefor strongly recommend to create your own account:

Petitioner Merck, Ex. 1066, p. 13

---

1. How do I work with SWISS-MODEL workspace ? => How can I manage my projects?

In the workspace a list of the current modeling work units is displayed, including the workunit type, a title provided by the user, and the status of the workunit:

Workunit	Type	Title	Status
P000001	Template Identification	Cyclin A1	
P000002	Modelling Automatic	Cyclin A1 (cyclin domain)	
P000003	Modelling Automatic	Cyclin A1 Model based on 1finB	
P000004	Sequence Scanning	Cyclin A1 - SS prediction	
P000005	Structure Assessment	Template 1finB for Cyclin A1	
P000006	Structure Assessment	Cyclin A1 Model based on 1finB	

The current status of a work unit is indicated by a graphical symbol: *submitted* (the job has been submitted to the queuing system and is waiting for execution), *running* (job is currently running and programs are calculating), *finished* (job has been completed, results are available) or *failed/stopped* (something went wrong during the process).

Symbols:
- submission not finished
- queued
- running
- failed/stopped
- finished

After completion of the modelling procedure (~ a few minutes up to several hours), the results are stored in the workspace and the user is notified about the completion.The user can access the results output by clicking on the work unit ID number. The results are stored for one week on the

Petitioner Merck, Ex. 1066, p. 14

server. The remainig time before deletion of a given work unit is also displayed. The user can decide to either delete a work unit or to prolonge its life span by clicking on the corresponding link.

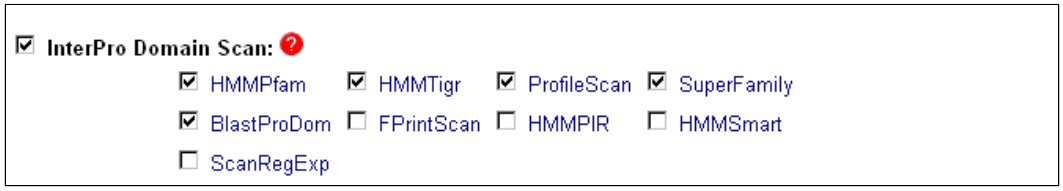
Beware: Workunits are kept on the server for one week before they are deleted automatically. You may postpone deletion by one week by pressing the green "refresh arrow". Please download the modelling results within this timeframe to your local system. Each user has a quota of up to a maximum of 25 work units which can be stored simultaneously.

---

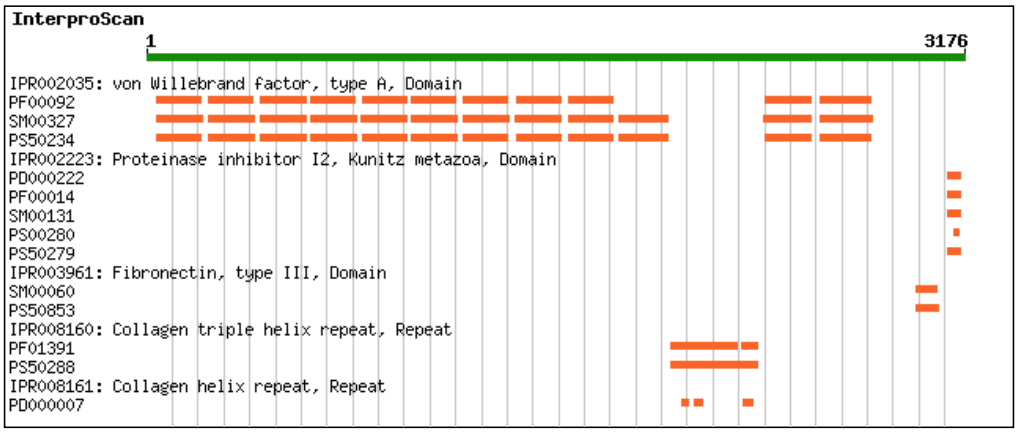
2. My protein is quite large, and would like to identify individual domains I could model separately.

Many proteins are modular and made up of several structurally distinct domains, which often reflect evolutionary relationships and may correspond to units of molecular function.The sensitivity and performance of profile-based template search methods can often be improved when the template search is performed on individual domains rather than the whole target sequence. The member databases of InterPro (*Mulder et al.*) allow for both the identification of protein domains and the assignment of protein function. Using the InterPro Domain Scan (IprScan, *Zdobnov et al.*), protein domains and functional sites can be assigned to regions of a target sequence.

See: [Tools] [Secondary Structure Prediction and Domain Assignment]



Let's use the example of Collagen alpha 3(VI) chain (UniProt accession code: **P12111**) to identify individual domains in the target sequence. The result looks like this:



The location of the individual domains is provided in tabular form below the graphics. Links to the motif definition in InterPro are provided.

```
Interpro Scan has finished. Here are the results:

IPR002035: von Willebrand factor, type A, Domain
  PF00092: 39 - 213 VWA
  PF00092: 242 - 415    VWA
  PF00092: 445 - 620    VWA
  PF00092: 639 - 812    VWA
  PF00092: 837 - 1009   VWA
  PF00092: 1029 - 1201  VWA
  PF00092: 1233 - 1404  VWA

  PF00092: 1436 - 1609  VWA
  PF00092: 1639 - 1812  VWA
  PF00092: 2402 - 2581  VWA
  PF00092: 2619 - 2810  VWA
```

Petitioner Merck, Ex. 1066, p. 15

If I have no idea about possible templates for my target, and I want to identify possible template structures.

The degree of difficulty in identifying a suitable template for a target sequence can range from "trivial" for well-characterized protein families to "impossible" for proteins with an unknown fold. The SWISS-MODEL Workspace provides access to a set of increasingly complex and computationally demanding methods to search for templates within the SWISS-MODEL Template library.

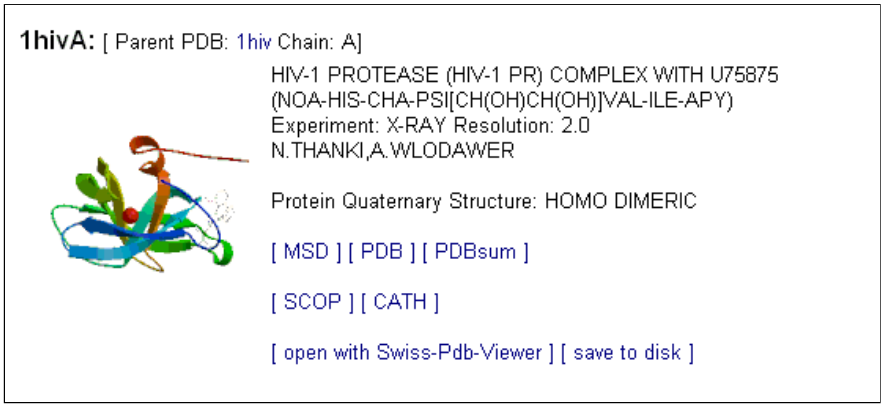
SwissModel Template Library (ExPDB)

The template structure database used by SWISS-MODEL (SMTL or ExPDB library) is derived from the Protein Data Bank (*Westbrook et al.*). In order to allow sequence-based template searches, each PDB entry is split into individual chains. The separated template chains are annotated with information about experimental method, resolution (if applicable), ANOLEA mean force potential (*Melo et al.*), Gromos96 energy (*van Gunsteren et al.*) and PQS (*Henrick et al.*) quaternary state assignment to allow for rapid retrieval of the relevant structural information during template selection. Theoretical models, structures only consisting of C alpha atoms and irregularly formatted database entries are removed. Templates sharing 100% sequence identity are grouped into a SMTL100 library using the program CD-HIT (*Li et al.*). Clusters of sequences having 90%, 70% and 50% sequence identity are derived from the RCSB non-redundant PDB lists.

[Tools] [SwissModel Template Library]

You may query if a certain PDB entry is part of SMTL. In this example, we search for chains of PDB entry "1HIV". SMTL provides information about the experimental methods used for structure determination, resolution (if applicable), and links to the original PDB entry as well as protein structure classification by SCOP and CATH.

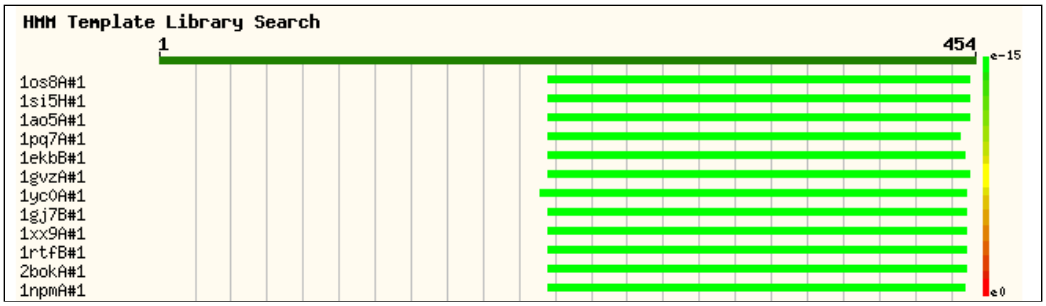
Caveat: A significant part of proteins are multimeric in their biologically active state. Single chains, or raw PDB entries often do not represent the biologically correct assembly. The PQS Protein Quaternary Structure Server (Henrick et al.) allows for searching of the list of likely quaternary structures generated at the EBI. As in our example, HIV-I protease is known to be active as a dimer. Multimeric proteins can be modelled in SWISS-MODEL Workspace using the **Project Mode.**



The target sequence can be used to query the SMTL for suitable template structures using "Template identification" in the Tools menu:

[Tools] [Template Identification]

A condensed graphical view of the modeling task is provided containing the target sequence, the template matches sorted and colored according to the associated E-value. Clickable bars indicate the matched regions and guide the user to the underlying original program output.



Petitioner Merck, Ex. 1066, p. 16

```
 Template: 1os8A  Length: 235   Eval: 8.51e-65
[Show Template Cluster 1os8A]
 [Display Alignment in DeepView]


 1os8A    1 VVGGTRAAQGEFPFMVRLS....MGCGGALYAQDIVLTAAHCVSGSGNNTSI 48
              VGG        P    L       CGG         TAAHCV       S
Target  217 IVGGNMSLLSQWPWQASLQfqgyHLCGGSVITPLWIITAAHCVYDLYLPKSW 268

 1os8A   49 TATGGVVDLQSSSAVKVRSTKVLQAPGYNG..TGKDWALIKLAQPINQP..T 96
            T   G V L    A      K      Y     G D AL KLA P
Target  269 TIQVGLVSLLDNPAPSHLVEKIVYHSKYKPkrLGNDIALMKLAGPLTFNemI 320
```

[Display Alignment in DeepView]

Target-template alignments from the search tools (BLAST or SAM) can be visualized in DeepView to correct misplaced insertions and deletions in the structural context of the template, and to manually adjust misaligned regions. The modified project can then be saved to disk and submitted as "project mode" to the workspace for model building by the SWISS-MODEL pipeline.

How do I use the fully automatic mode of SWISS-MODEL workspace?

The "automated mode" is suited for cases where the target-template similarity is sufficiently high to allow for fully automated modelling. As a rule of thumb, automated sequence alignments are sufficiently reliable when target and template share more than 50% percent of sequence identity.

This submission requires only the amino acid sequence (FASTA format or single letter raw sequence) or the UniProt accession code of the target protein as input data. The modelling pipeline automatically selects suitable templates based on a Blast E-value limit, which can be adjusted upon submission (*Altschul et al.*). The automated template selection will favour high-resolution template structures with reasonable stereochemical properties as assessed by ANOLEA mean force potential (*Melo et al.*) and Gromos96 force field energy (*van Gunsteren et al.*).

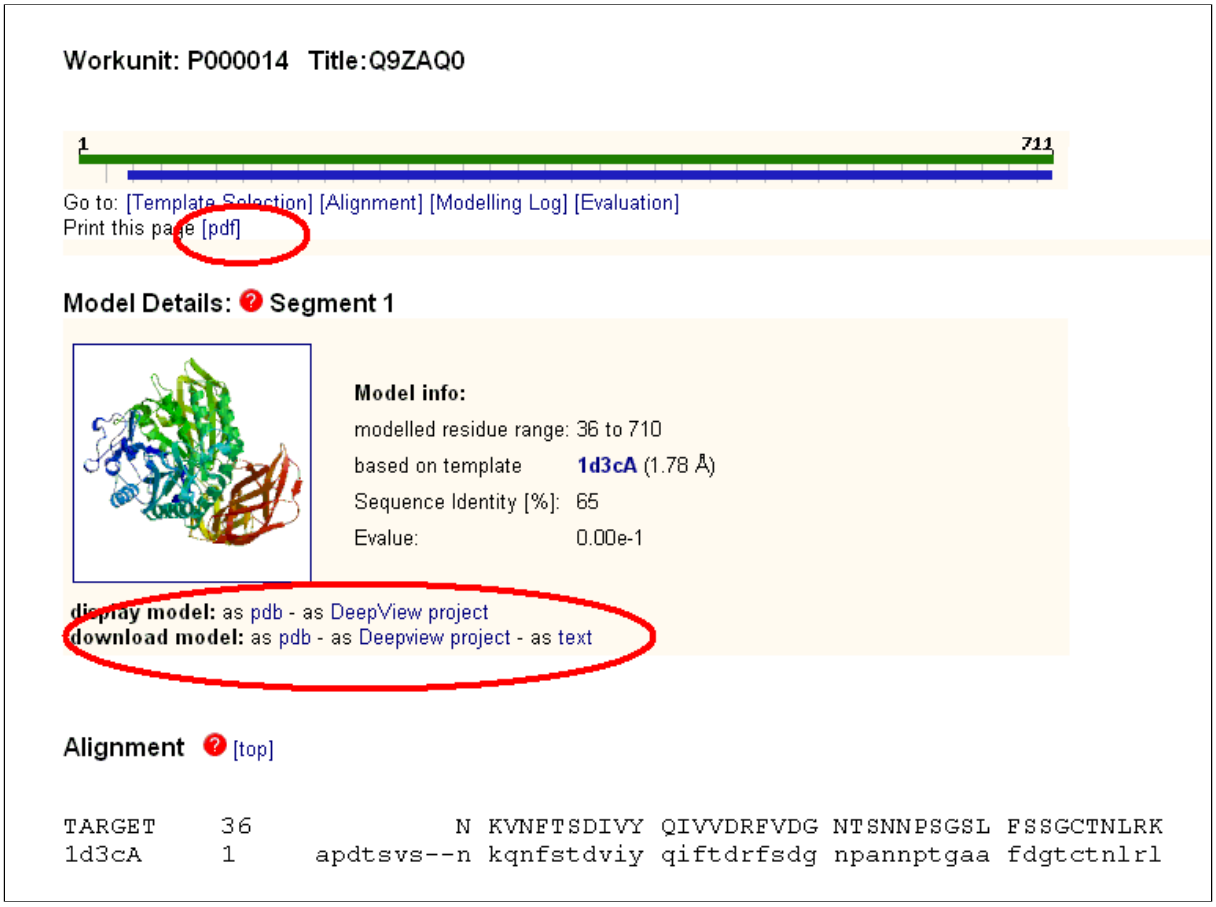
Example: Modelling the catalytic domain of Cyclodextrin glucanotransferase from Bacillus stearothermophilus (UniProt AC code: Q9ZAQ0).

[Modelling] [Automated Mode]

Note: Workunits will be automatically deleted after 1 week from the server. When the modelling project is finished, please download the results and save them locally:

- **Model coordinate file** (PDB format or DeepView project format) PDB files can be displayed using DeepView, Dino, or otehr tools for molecular visualization.
- **Logfile PDF**: The content of the web page (including images and logfiles, but not the model coordinates) can be downloaded and saved as PDF. See "Print this page [pdf]" at the top of the page. PDF files can be displayed using Acrobat Reader.

Petitioner Merck, Ex. 1066, p. 17



Alignment Mode

Multiple sequence alignments are a common tool in many molecular biology projects. If the three-dimensional structure is known for at least one of the members, this alignment can be used as starting point for comparative modelling using the "alignment mode".
The "alignment mode" allows the user to test several alternative alignments and evaluate the quality of the resulting models in order to achieve an optimal result.

In order to facilitate the use of alignments in different formats, the submission is implemented as a three step procedure:

1. Prepare a multiple sequence alignment.

- It must contain at least your target sequence and the template sequence
- Use any of your favorite alignment tools. We recommend T_COFFEE by Cedric Notredame
- Make sure the sequence names are "reasonable"

2. Submit your alignment to the Workspace Alignment Mode.

- Possible formats are: FASTA, MSF, CLUSTALW, PFAM and SELEX
- You may either upload your file or cut & paste
- Don't forget to specify the correct alignment format
- Here is a small example for testing (cut & paste):

```
CLUSTAL W (1.82) multiple sequence alignment

THN_DENCL       KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH- 46
THNX_TEST       KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS-YPNK 46
1crnA           TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN- 46
                .:***  ..*  :  **: * .. :**  :** **..: **  *
```

3. Select Target and Template

- The alignment (as it was interpreted by the server) should now be displayed in the bottom part of the page.
- The script will try to make a good guess for the correct names based on your submission.
- Select the sequence name of the target sequence (e.g. THN_DENCL)

Petitioner Merck, Ex. 1066, p. 18

- Select the sequence of the template structure (e.g. 1crnA). You don't need to use PDB IDs, you may use any name you like.
- Specify the template structure to which this sequence belongs. This template MUST be part of the ExPDB template library. Please use the SWISS-MODEL Template library tool to check...
- Don't forget to specify the correct CHAIN ID. Note that PDB's chain IDs are normally in capital letters.

```
Target   sequence: THN_DENCL
Template sequence: 1crnA        PDB-Code: 1crn      Chain-ID: _
```

4. Check Alignment and Submit

- The alignment at the bottom of the page should represent the correct mapping of the template structure on the target sequence. Please check carefully before submission.
- As usual, please provide name and e-mail for the SWISS-MODEL submission.
- Good Luck with you model ....

The server pipeline will build the model purely based on this alignment. During the modelling process, implemented as rigid fragment assembly in the SWISS-MODEL (*Schwede et al.*) pipeline, the modelling engine might introduce minor heuristic modifications to the placement of insertions and deletions.

Supported Alignment formats

The following formats are currently supported: FASTA, MSF, CLUSTALW, PFAM and SELEX;

Examples:

fasta:

```
>THN_DENCL
KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH-
>THNX_TEST
KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS-YPNK
>1crnA
TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN-
```

clustal:

```
CLUSTAL W (1.82) multiple sequence alignment

THN_DENCL       KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH- 46
THNX_TEST       KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS-YPNK 46
1crnA           TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN- 46
                .:***  ..*  :  **: * .. :**  :** **..: **  *
```

msf:

```
 !!AA_MULTIPLE_ALIGNMENT 1.0

  thn_dencl.msf MSF:  47 Type: P 08/08/05 CompCheck:  427 ..

  Name: THN_DENCL  Len: 47  Check: 8212 Weight: 1.00
  Name: THNX_TEST  Len: 47  Check: 5295 Weight: 1.00
  Name: 1crnA      Len: 47  Check: 6920 Weight: 1.00

//

           1                                           47
THN_DENCL  KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH~
THNX_TEST  KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS.YPNK
1crnA      TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN~
```

How do I use the Project Mode mode of SWISS-MODEL workspace ?

Main application: ***Visual inspection of alignments; modelling of Oligomeric proteins.***

In difficult modeling situations, where the correct alignment between target and template cannot be clearly determined by sequence based methods, visual inspection and manual manipulation of the alignment can significantly help improving the quality of the resulting model. Project files containing the superposed template structures, and the alignment between the target and the template can be generated using the program

Petitioner Merck, Ex. 1066, p. 19

DeepView (*Swiss-PdbViewer Guex et al*).The user has therfor full control over essential modelling parameters, i.e. the choice of template structures, the correct alignment of residues, and the placement of insertions and deletions in the context of the three-dimensional structure. Modelling of oligomeric proteins with Swiss-Model Workspace can be done using the Project Mode.

The program DeepView can be downloaded freely from the *ExPASy* web site. DeepView does not require administrator privileges for installation. E.g. under MS windows, simply uncompress the distributed archive at any location you like (e.g. c:\spdbv or on your desktop) and start working by starting the spdbv.exe application. Tutorials, manuals and discussion group for DeepView can be found on the DeepView web site.

Example: Modelling a dimeric protein

In order to demonstrate Oligomer-Modelling, we are going to build a model of the protease of murine leukemia virus based on thestructure of Nelfinavir-resistant HIV-1 protease (D30N/N88D) in complex with Darunavir [3HVP]. (Please keep in mind that this just an example to illustrate the workflow. Most likely using this template will not make much scientific sense in most cases.)

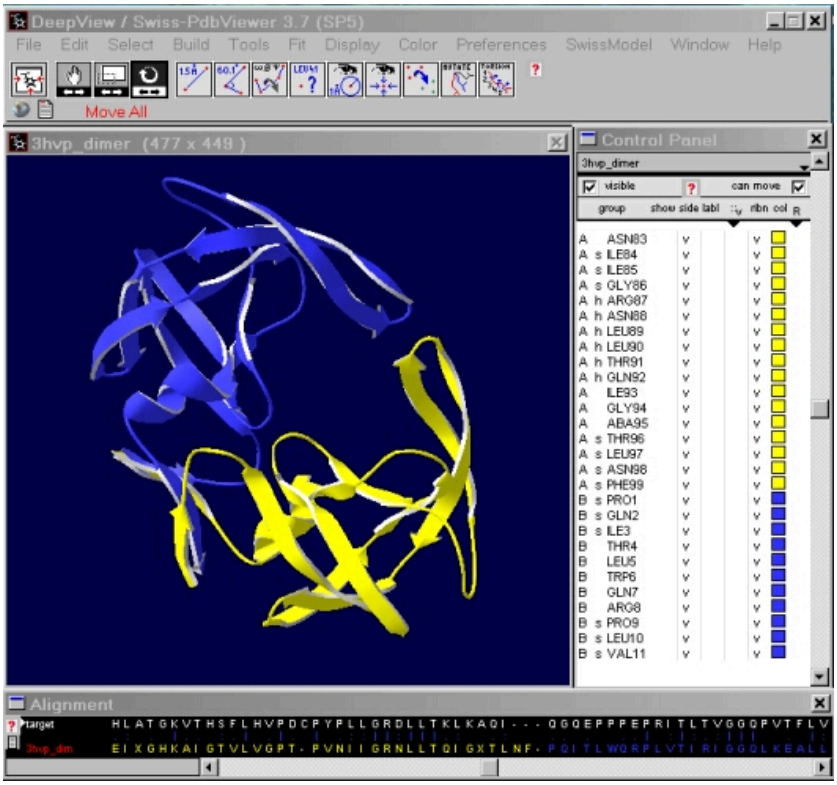
1. **Get the template in the correct quaternary state**
   First, check the correct biological assembly of your template protein. Copies of the assymetric unit of the PDB files can be generated by applying the correct crystallographic symmetry operators. The PDB download page will allow you to download a “biological assembly” file. If you are unsure which assembly to use, the [PISA] server helps to visualize alternative oligomeric states. Download and save the template coordinates as PDB file to your local disk. [3lzv_dimerAB.pdb].
2. **Remove all non-aminoacid residues**
   Open the file in DeepView and remove all non-aminoacid groups such as ions, ligands, OXT, etc. 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).
3. **Ensure Unique Chain IDs**
   Make sure each chain has a unique name, e.g. "A","B", etc. Coloring the molecule by chain helps to check. Here is an example file for download for this tutorial [3lzv_dimerAB.pdb].
4. **Target Sequence**
   In our example, we will model the protease domain of murine leukemia virus (UniProt AC: P03356). As you can see, the virus encoded polyprotein consists of several domains. Before modelling, it make things easier to focus on the interesting segment. You may use e.g. the IprScan utility to identify the individual domains. In our case, we will use residue 3-100.

   Create a FASTA file with your target sequences for each chain in the SAME order as in the template, i.e. "A", then "B" etc separated by semicolons. [target.txt]

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

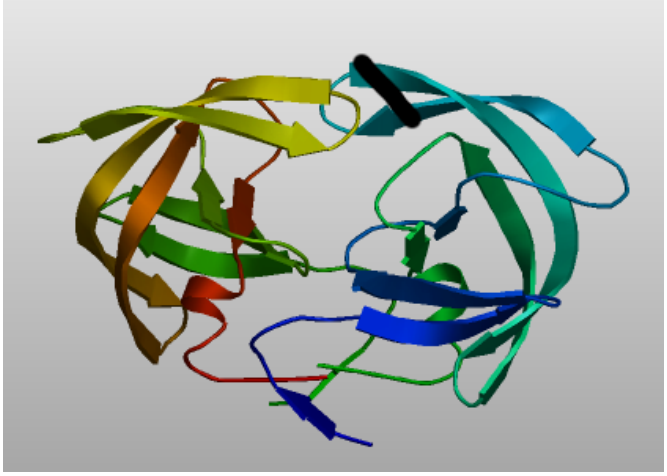
5. **Load the target sequence into DeepView**
   Please make sure to start with loading the amino acid sequence of your target protein *first* using the "SWISS-MODEL" menu - before loading any template structures.
6. **Load the template structure into DeepView**
   and generate a preliminary target-template alignment using Menu: Fit - Fit raw sequence.
7. **Adjust target-template Alignment in DeepView**
   Open the alignment window and adjust alignment. Make sure NOT to align residues of different chains ("color by chains" helps to see the chain boundaries in both sequences). Do not align to "non aminoacid residues" like het groups, OXT. Make sure all insertions & deletions are correctly positioned in the structural context.

Petitioner Merck, Ex. 1066, p. 20



8. **SWISS-MODEL Submission**
   Save the project to your local disk [e.g. tutorial_dimerAB.pdb] and submit the file to the project mode of SWISS-MODEL workspace for model building. A new workunit will be created, containing the modelling results, includung log file, ANOLEA evaluation, and model project file of the modelled dimer.

[Modelling] [Project Mode]



Model of the dimeric protease.

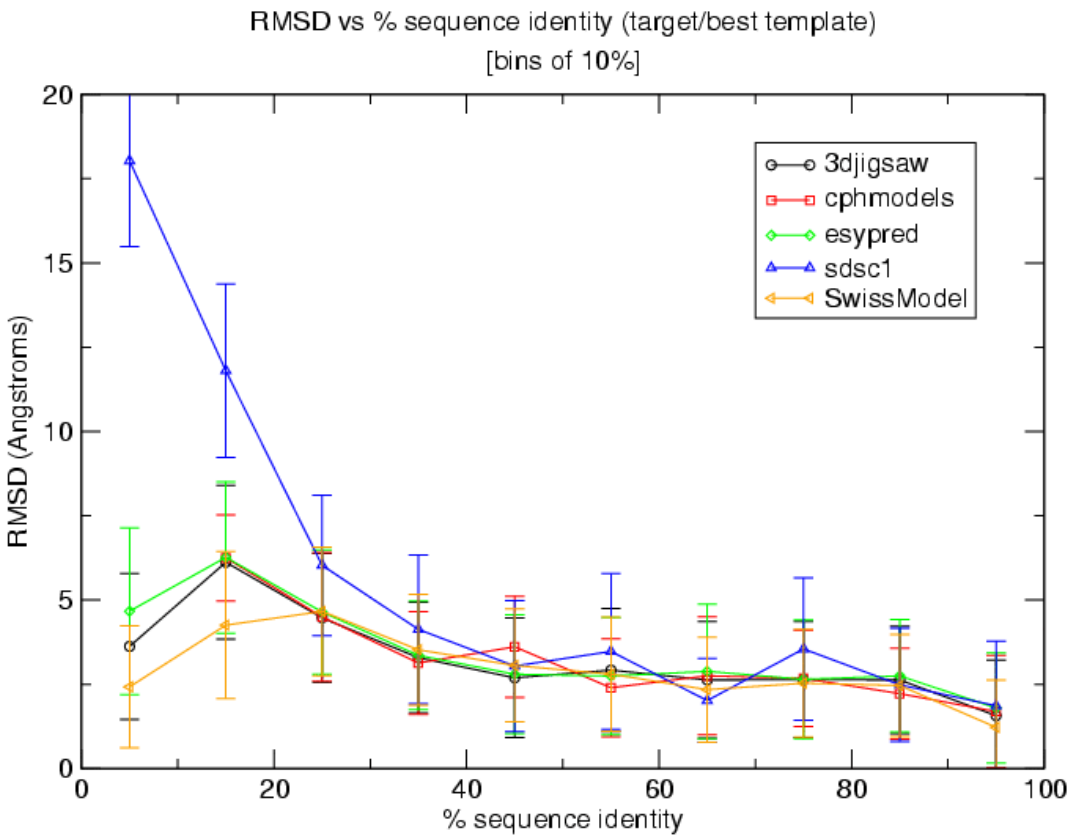
What accuracy can I expect for a model build by the automated mode of SWISS-MODEL?

Evaluation of template structure and model quality is a crucial step in homology modelling. The reliability of different protein modeling methods can be assessed by evaluating the results of blind predictions after the corresponding protein structures have been determined experimentally. The overall performance of the SWISS-MODEL pipeline is evaluated by the EVA project. SWISS-MODEL was the first comparative modelling server to join the EVA project in May 2000, and has since then been continuously evaluated. As of Summer 2005, EVA-CM is based on the assessment of 261 weekly releases of the PDB database, resulting in 48098 protein models for 19698 protein target chains for five different prediction servers, among these 18314 from SWISS-MODEL. All models generated by SWISS-MODEL server, evaluation results, score definitions and detailed statistics are available from the EVA project website.

The C-alpha atoms RMSD after global superimposition of the model and the experimental target structures was computed and plotted vs. % of sequence identity between target and best template to give an estimation of the overall accuracy of the different modelling servers with regards to

Petitioner Merck, Ex. 1066, p. 21

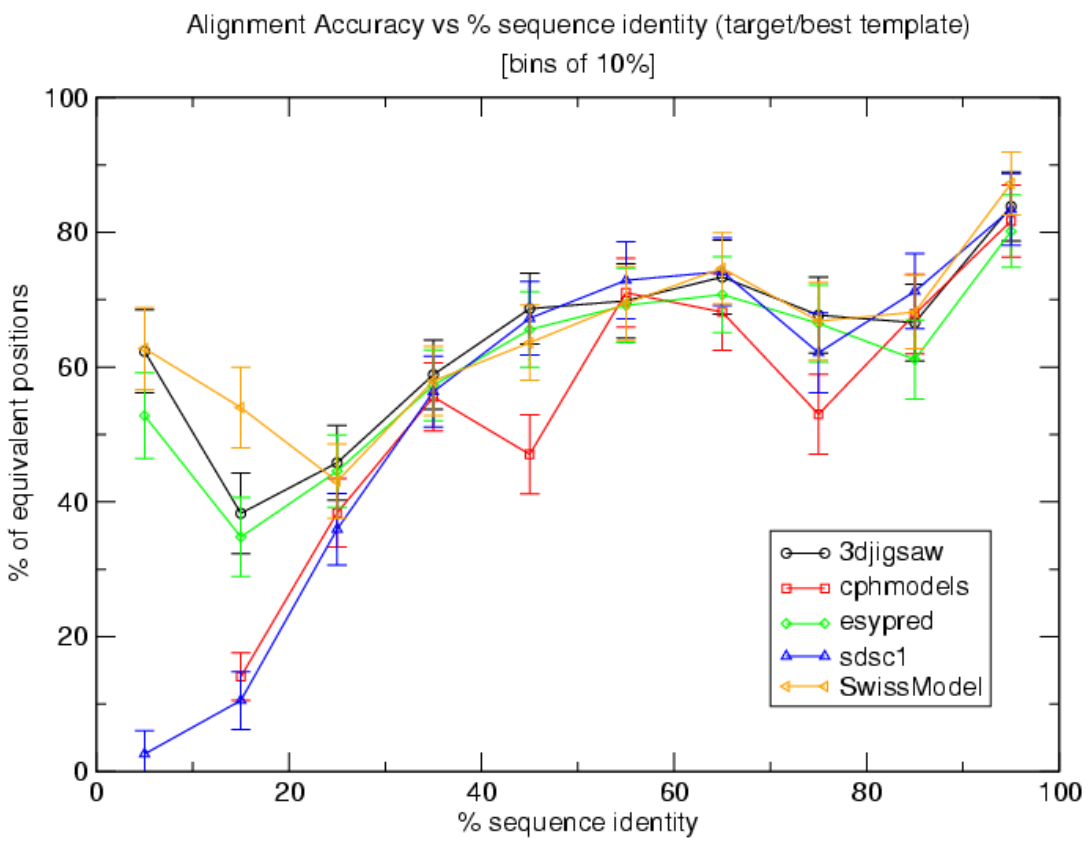
different sequence identities between target and template:



In general, major differences between the individual prediction methods are only observed for target-template pairs sharing sequence identities of less than 40 %, where methods favouring higher coverage of the target sequences are more likely to generate models with a higher RMSD. As expected, model RMSD is increasing with decreasing alignment accuracy as defined by the percentage of equivalent C-alpha positions (within 3.5

Petitioner Merck, Ex. 1066, p. 22

Angstroms) between the optimally superimposed target and model structures:



How can I assess a structure or model with empirical force-field and Mean Force Potential methods?

Evaluation of model quality is a crucial step in homology modeling. While the performance of the automated SWISS-MODEL (*Schwede et al.*) pipeline in general is continuously evaluated by the EVA project (*Koh et al.*), the quality of individual models can vary significantly.

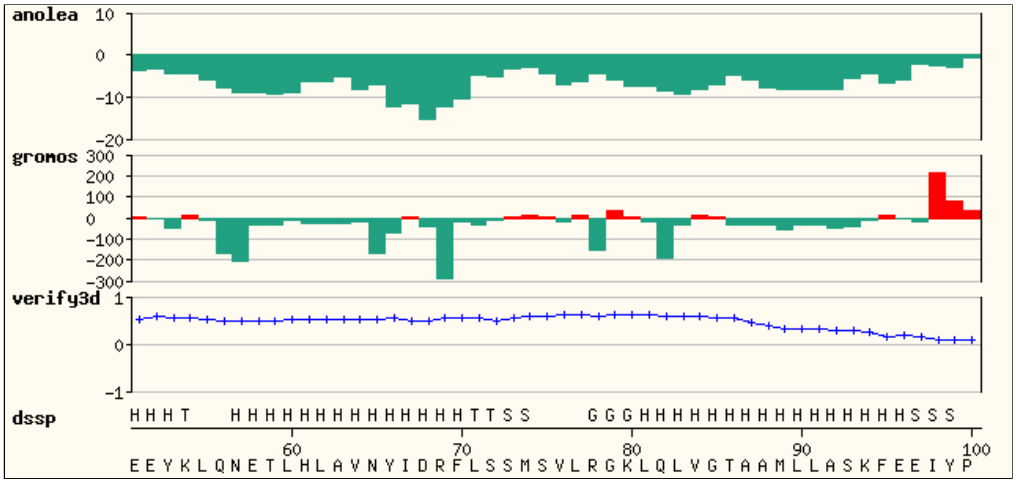
Therefore, graphical plots of Anolea mean force potential (*Melo et al.*), GROMOS empirical force field energy (*van Gunsteren et al.*) and Verify3D profile evaluation (*Eisenberg et al.*) are provided to enable the user to estimate the quality of protein models and template structures.

Anolea: The atomic empirical mean force potential ANOLEA (*Melo et al.*) is used to assess packing quality of the models. The program performs energy calculations on a protein chain, evaluating the "Non- Local Environment" (NLE) of each heavy atom in the molecule. The y-axis of the plot represents the energy for each amino acid of the protein chain. Negative energy values (in green) represent favourable energy environment whereas positive values (in red) unfavourable energy environment for a given amino acid.

Verify3D: The Verify3D (*Eisenberg et al.*) method assess protein structures using three-dimensional profiles. This program analyzes the compatibilty of an atomic model (3D) with its own amino acid sequence (1D). Each residue is assigned a structural class based on its location and environment (alpha, beta, loop, polar, apolar etc). Then a database generated from good structures is used to obtain a score for each of the 20 amino acids in this structural class. The vertical axis in the plot represents the average 3D-1D profile score for each residues in a 21-residue sliding window. The scores ranges from -1 (bad score) to +1 (good score).

Gromos: The y-axis of the plot represents the GROMOS (*van Gunsteren et al.*) empirical force field energy for each amino acid of the protein chain. Negative energy values (in green) represent favourable energy environment whereas positive values (in red) unfavourable energy environment for a given amino acid.

Petitioner Merck, Ex. 1066, p. 23



How can I assess geomoetrical accuracy of a structure or model?

Evaluation of model quality is a crucial step in homology modeling. While the performance of the automated SWISS-MODEL (*Schwede et al.*) pipeline in general is continuously evaluated by the EVA project (*Koh et al.*), the quality of individual models can vary significantly.

Therefore, graphical plots of Anolea mean force potential (*Melo et al.*), GROMOS empirical force field energy (*van Gunsteren et al.*), Verify3D profile evaluation (*Eisenberg et al.*), Whatcheck (*Hooft et al.*) and Procheck (*Laskowski et al.*) reports are provided to enable the user to estimate the quality of protein models and template structures.

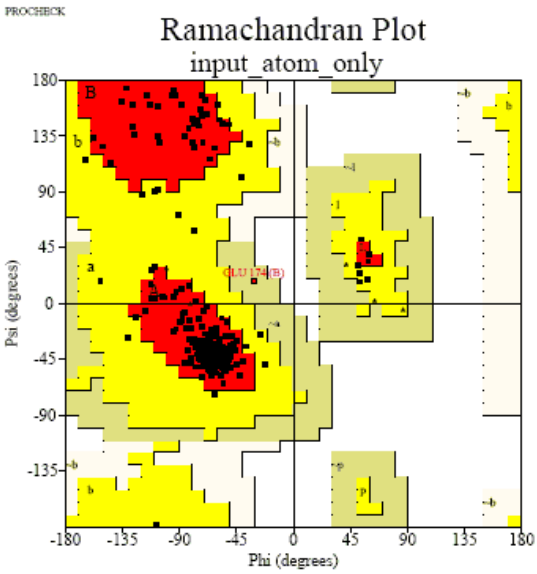
Procheck

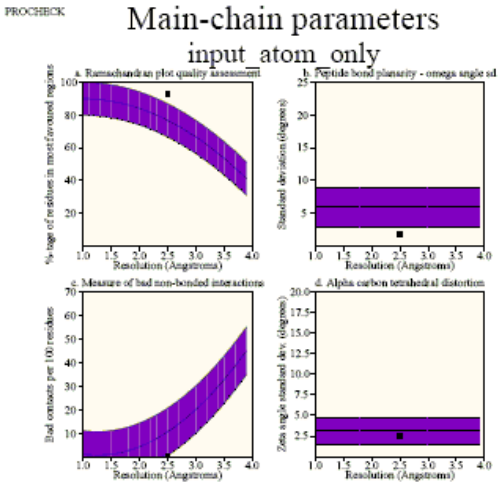
The PROCHECK suite of programs (*Laskowski et al.*) assess the "stereochemical quality" of a given protein structure. The aim of PROCHECK is to assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures.

What Check

What Check comprises several tools for protein structure verification (*Hooft et al.*). Besides of a detailed report, a summary for "users of a structure" is provided. Detailled documentation for the WHAT_CHECK output is available at the WHAT_CHECK homepage.

Example outputs:





SWISS-MODEL is developed by the Protein Structure Bioinformatics group at the SIB - Swiss Institute of Bioinformatics & the Biozentrum University of Basel. © 2011.

Petitioner Merck, Ex. 1066, p. 24

The Wayback Machine - https://web.archive.org/web/20110519141354/http://swissmodel.expasy.org/workspace/index.php?func=modelling_si…

SWISS-MODEL Workspace

[myWorkspace] [login]

SwissModel Automatic Modelling Mode

Email:

Project Title:

Provide a protein sequence or a UniProt AC Code:

Submit Modelling Request

Advanced options:

Use a specific template: PDB-ID: Chain:

or

Template file: Choose File No file chosen

Petitioner Merck, Ex. 1066, p. 25

SWISS-MODEL is developed by the Protein Structure Bioinformatics group at the SIB - Swiss Institute of Bioinformatics & the Biozentrum University of Basel. © 2011.

Petitioner Merck, Ex. 1066, p. 26

The Wayback Machine - https://web.archive.org/web/20110519144444/http://swissmodel.expasy.org/workspace/index.php?func=tools_structur…

SWISS-MODEL Workspace

[myWorkspace] [login]

Protein Structure & Model Assessment Tools

Email:
Project Title:

The following tools are provided to assess the quality and structural features of protein models and template structures.
Please upload a model or template structure in **PDB format**.

Choose File No file chosen

submit

Local Model Quality Estimation:

- [x] Anolea — Anolea atomic mean force potential
- [ ] Gromos — Empirical force field
- [x] QMEAN6 — Composite scoring function for model quality estimation [Example]

Global Model Quality Estimation:

- [x] DFire — All-atom distance-dependent statistical potential

Stereochemistry Check:

- [ ] Procheck — Stereochemical quality check; min. Resolution: 2.5 Å

Structural Features:

- [ ] DSSP — Secondary Structure, geometrical features, and solvent exposure assignment
- [ ] Promotif — Analysis of protein structure motifs

* Disclaimer and user agreement: The use of SWISS-MODEL Server and SWISS-MODEL Workspace is free for both academic and commercial users. However, some of the tools provided on this page may require a valid license for some users. If you are in doubt, please check carefully the licensing conditions for each tools. Links to the original sites are provided in the help file. By using SWISS-MODEL workspace, you confirm that you are legally entitled to use these tools. You also agree to cite SWISS-MODEL Workspace AND the corresponding reference of these tools in any of your publications reporting results obtained from this service.

Petitioner Merck, Ex. 1066, p. 27

Swiss Institute of Bioinformatics | About SWISS-MODEL | Privacy | Terms of use | News Back to the Top

SWISS-MODEL is developed by the Protein Structure Bioinformatics group at the SIB - Swiss Institute of Bioinformatics & the Biozentrum University of Basel. © 2011.

Petitioner Merck, Ex. 1066, p. 28

SWISS-MODEL Template Library

[close]

Introduction to SWISS-MODEL Workspace

The SWISS-MODEL Workspace is a web-based integrated service dedicated to protein structure homology modelling. It assists and guides the user in building protein homology models at different levels of complexity.

Building a homology model comprises four main steps: identification of structural template(s), alignment of target sequence and template structure(s), model building, and model quality evaluation. These steps can be repeated until a satisfying modelling result is achieved. Each of the four steps requires specialized software and access to up-to-date protein sequence and structure databases.

Protein sequence and structure databases necessary for modelling are accessible from the workspace and are updated in regular intervals. Software tools for *template selection*, *model building*, and structure *quality evaluation* can be invoked from within the workspace.

A personal working environment (*workspace*), where several modelling projects can be carried out in parallel, is provided for each user.

This help file provides references and illustrate the use of the individuals tools available from within the SWISS-MODEL Workspace.
A tutorial to facilitate the first steps of working with SWISS-MODEL Workspace as a list of most frequently asked questions is provided here: Tutorial
Please also take a look at the following published [Protocol]

Workspace

The SWISS-MODEL Workspace provides a personal web-based area for each user in which protein homology models can be built and the results of completed modelling projects are stored and visualized.

In the workspace a list of the current modeling work units and their current status is displayed: *submitted* (the job has been submitted to the pipeline but still queuing), *running* (job is running and programs are calculating), *finished* (job has been completed, final results are available) or *failed/stopped* (if something went wrong during the process).

Depending on the type of job the user has submitted a different tag will be associated with a work unit: *Template Identification* for template identification, *Sequence Scanning* for secondary structure and disorder prediction and domain assignment, *Structure Assessment* for structure quality assessement. And *Modelling Automatic*, *Modelling Project*, *Modelling Alignment* respectively for automated, alignemnt or project mode modeling requests.

After completion of the modelling procedure (~ a few minutes up to several hours), the results are stored in the workspace and the user is notified about the completion.The user can access the results output by clicking on the work unit ID number.

The results are stored for one week on the server. The remainig time before deletion of a given work unit is also displayed. The user can decide to either delete a work unit or to prolonge its life span by clicking on the corresponding link.

Beware: Each user can submit up to a maximum of 25 work units.

Domain assignment, Secondary Structure and Disorder Prediction

Many proteins are modular and made up of several structurally distinct domains, which often reflect evolutionary relationships and may correspond to units of molecular function.The sensitivity and performance of profile-based template search methods can often be improved when the template search is performed on individual domains rather than the whole target sequence. IprScan (see below) allows for protein domains and functional site prediction.
Protein disorder prediction measures and displays the propensity of protein sequences to be ordered or disordered. The result can aid the assignment of templates to a specific region of the target protein by complementing the IprScan approach to globular domains and feature discovery.
Secondary structure prediction methods are especially useful when combined with other types of analyses: e.g. in cases where only templates with very low sequence homology can be detected by sequence-based search methods, predicted secondary structure may help to decide if a putative template shares structural features of the target protein.

InterPro Domain Scan

The member databases of InterPro (*Mulder et al.*) allow for both the identification of protein domains and the assignment of protein function. Using the InterPro Domain Scan (IprScan, *Zdobnov et al.*), protein domains and functional sites can be assigned to regions of a target sequence.

The following databases are currently part of the InterPro Domain scan method:

HMMPfam: Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families.

HMMTigr: TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology.

ProfileScan: PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs. There are a number of protein families as well as functional or structural domains that cannot be detected using patterns (see below) due to their extreme sequence divergence. The use of techniques based on weight matrices (also known as profiles) allows the detection of such domains.

SuperFamily: SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure, based on SCOP.

BlastProDom: The ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches. The ProDom database has been designed as a tool to help analyze domain arrangements of proteins and protein families.

FPrintScan: PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family.

HMMSmart:SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures.

ScanRegExp:PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs. Some biologically significant amino acid patterns can be summarised in the form of regular expressions.

The results of the InterPro domain mapping is displayed in combination with the alignment to putative template structures, allowing the user to identify template structures spanning one or more domains of the target protein. For low homology templates, the IprScan functional site annotation of the target sequence can be used to verify that putative templates share essential functional features. The InterPro functional annotations for individual template structures are accessible from the workspace as links to the SMTL library and external resources..

PsiPred Secondary Structure Prediction

PSIPRED is a method for protein secondary structure prediction (*Jones DT et al*.).

The plot shows position in the sequence against probability of being be part of a alpha helix (H) , extended beta strand (E) or a coil region (C). The result of the prediction is plottet on the x-axis of the plot.

DISOPRED Disorder Prediction

DISOPRED (v 2) is a neural-network based predictor of disordered regions in proteins (*Jones DT et al*.).

The majority of water-soluble proteins have structures that are globular and relatively static. However, some proteins have regions that are natively disordered. Disordered regions are flexible, dynamic and can be partially or completely extended in solution. Native disorder also exists in global structures such as extended random coil proteins with negligible secondary structure or molten globules, which have regular secondary structure elements but have not condensed into a stable globular fold.The primary function of disorder appears to be molecular recognition of proteins and nucleic acids. It has been speculated that the multiple metastable conformations, adopted by disordered binding sites, allows recognition of several targets with high specificity and low affinity. Order to disorder transitions also provide a mechanism for controlling protein concentration via proteolytic degradation.

The plot shows position in the sequence against probability of being disordered (from 0 to 1). The 'filter' curve represents the outputs from DISOPRED and the 'output' curve the outputs from a linear SVM classifier (DISOPREDsvm). The outputs from DISOPREDsvm are included to indicate shorter, low confidence predictions of disorder.

Asterisks (*) represent disordered predictions and dots (.) prediction of order.

The disopred predictions are given at a default false positive rate threshold of 2%. But this value can be changed by the user.

MEMSAT

MEMSAT 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.

Template Identification

The degree of difficulty in identifying a suitable template for a target sequence can range from "trivial" for well-characterized protein families to "impossible" for proteins with an unknown fold. The SWISS-MODEL Workspace provides access to a set of increasingly complex and computationally demanding methods to search for templates.

Templates which are close homologues of the target can be identified using a gapped BLAST (*Altschul et al.*) query against the ExPDB template library extracted from PDB.

Petitioner Merck, Ex. 1066, p. 30

Options for the **BLAST** database search are:
E-value cutoff: sets the threshold expectation value for keeping alignments. It describes how often a given score is expected to occur random;
Matrix: the protein substitution matrix;
SEG Filter: filters the query sequence for low-complexity subsequences;
Descriptions: sets the number of database sequences for which to show the one-line summary descriptions at the top of a BLAST report;
Alignments: truncates the report to the selected number of alignments;

When no suitable templates are identified, or only parts of the target sequence are covered, two additional approaches for the sensitive detection of distant relationships among protein families are provided:

Iterative Profile Blast: the template library is searched with PSI-BLAST (*Altschul et al.*) using an iteratively generated sequence profile based on NR (*Wheeler et al.*). This method has been initially introduced as PDB-Blast by Godzik and coworkers.

- The first run searches the NR database and derive a profile for the query sequence. The following options are available:
Iterations: number of iteration for the NR database search and profile (PSSM) generation;
Matrix: the protein substitution matrix;
Evalue: The E-value threshold for inclusion in PSSM. All alignments better than this threshold are used in constructing the PSSM;
SEG Filter: filters the query sequence for low-complexity subsequences;

- Then with this profile, the final run searches the SWISS-MODEL template library (ExPDB). The following options are available:
Database to search: Clustered versions of ExPDB (e.g. ExPDB90, sequences clustered to 90% of redundancy) which combine closely related sequences into a single record;
E-value cutoff: sets the threshold expectation value for keeping alignments. It describes how often a given score is expected to occur random;
Matrix: the protein substitution matrix;
SEG Filter: filters the query sequence for low-complexity subsequences;
Descriptions: sets the number of database sequences for which to show the one-line summary descriptions at the top of a BLAST report;
Alignments: truncates the report to the selected number of alignments;

HHSearch: To detect distantly related template structures, a target sequence can be searched against a Hidden Markov Model (HMM) based template library. Each HMM of the library is based on a multiple sequence alignment of the template sequence built by PSI-BLAST search (against nr90 & nr70) enriched with secondary structure assignment.
In analogy a HMM is built for the target sequence, which is subsequently used to search against the template library. Only alignments which score more than a given P-value cut-off are reported. Model building and library searches are performed using the HHSEARCH (v. 1.5.01) software package (*Söding et al.*) with default parameters.
For detailed documentation, please visit the official HHSEARCH site [http://toolkit.tuebingen.mpg.de/hhpred]

Display of template identification results

A condensed graphical view of the modeling task is provided containing the target sequence, the template matches sorted and colored according to the associated E-value, and the InterPro mappings. Clickable bars indicate the matched regions and guide the user to the underlying original program output.

In the InterPro output a link leads to the detailed InterPro page for this entry.

In the output of the different template identification programs the template annotations (via the link to the SWISS-MODEL Template library) and target-template alignment can be retrieved.

Alignments can be obtained as DeepView project file. The latter allows the user to visualize the different alignments in the structural context of the template, to correct misplaced insertions and deletions, and to manually adjust misaligned regions. The modified project can then be saved to disk and submitted as "project mode" to the workspace for model building by the SWISS-MODEL pipeline.

When searching a clustered version of the SWISS-MODEL Template library (e.g. ExPDB90) only the alignment between the target sequence and the sequence of the representative of the cluster is shown. Information about the members of the cluster is presented in the detailed output of the different template search programs. For each template, the SWISS-MODEL workspace provides a summary showing a small ribbon representation, experimental details, information about bound molecules, as well as links to PDB (*Westbrook et al.*), SCOP (*Andreeva et al.*), CATH (*Pearl et al.*), PDBsum (*Laskowskiet et al.*), and MSD (*Velankaret et al.*).

Model building

Depending on the difficulty of the modelling task, three different types of modelling requests (*automated mode*, *alignment mode*, *project mode*) are provided, which differ in the amount of user intervention.

Modelling requests are computed by the SWISS-MODEL server homology modelling pipeline (*Schwede et al.*).

Automated Mode

The "automated mode" is suited for cases where the target-template similarity is sufficiently high to allow for fully automated modelling. As a rule of thumb, automated sequence alignments are sufficiently reliable when target and template share more than 50% percent of sequence identity.

This submission requires only the amino acid sequence or the UniProt accession code of the target protein as input data. The pipeline will automatically select suitable templates based on a Blast (*Altschul et al.*) E-value limit (which can be adjusted upon submission), experimental quality, bound substrate molecules, or different conformational states of the template.

Depending on the planned model application, it can be necessary to select a different structural template than the one ranked first in the automated process. Typical examples are proteins in different conformational states, e.g. 1ake vs. 4ake. It is possible to specify the structure to be used as modelling template either by identifying an entry in the SWISS-MODEL template library by PDB-ID + ChainID e.g. "1ake" chain "A", or by uploading a file in PDB format (*) with coordinates of the template structure. Please make sure that this file contains only a single protein chain, and does not contain chemically modified amino acids, hereto atoms, ligands, etc.

Petitioner Merck, Ex. 1066, p. 31

(*) A simple PDB-like file containing the coordinates of the template structure. For more information about PDB file format please see link: http://www.wwpdb.org/docs.html

Alignment Mode

Multiple sequence alignments are a common tool in many molecular biology projects. If the three-dimensional structure is known for at least one of the members, this alignment can be used as starting point for comparative modelling using the "alignment mode".
The "alignment mode" allows the user to test several alternative alignments and evaluate the quality of the resulting models in order to achieve an optimal result.

In order to facilitate the use of alignments in different formats, the submission is implemented as a three step procedure:

1. Prepare a multiple sequence alignment.

- It must contain at least your target sequence and the template sequence
- Use any of your favorite alignment tools. We recommend T_COFFEE by Cedric Notredame
- Make sure the sequence names are "reasonable"

2. Submit your alignment to the Workspace Alignment Mode.

- Possible formats are: FASTA, MSF, CLUSTALW, PFAM and SELEX
- You may either upload your file or cut & paste
- Don't forget to specify the correct alignment format
- Here is a small example for testing (cut & paste):

```
CLUSTAL W (1.82) multiple sequence alignment

THN_DENCL       KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH- 46
THNX_TEST       KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS-YPNK 46
1crnA           TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN- 46
                .:***  ..*  :  **: * .. :**  :** **..: **  *
```

3. Select Target and Template

- The alignment (as it was interpreted by the server) should now be displayed in the bottom part of the page.
- The script will try to make a good guess for the correct names based on your submission.
- Select the sequence name of the target sequence (e.g. THN_DENCL)
- Select the sequence of the template structure (e.g. 1crnA). You don't need to use PDB IDs, you may use any name you like.
- Specify the template structure to which this sequence belongs. This template MUST be part of the ExPDB template library. Please use the SWISS-MODEL Template library tool to check...
- Don't forget to specify the correct CHAIN ID. Note that PDB's chain IDs are normally in capital letters.

```
Target   sequence: THN_DENCL
Template sequence: 1crnA      PDB-Code: 1crn      Chain-ID:  A
```

4. Check Alignment and Submit

- The alignment at the bottom of the page should represent the correct mapping of the template structure on the target sequence. Please check carefully before submission.
- As usual, please provide name and e-mail for the SWISS-MODEL submission.
- Good Luck with your model ....

The server pipeline will build the model purely based on this alignment. During the modelling process, implemented as rigid fragment assembly in the SWISS-MODEL (*Schwede et al.*) pipeline, the modelling engine might introduce minor heuristic modifications to the placement of insertions and deletions.

Supported Alignment formats

The following formats are currently supported: FASTA, MSF, CLUSTALW, PFAM and SELEX;

Examples:

fasta:

```
>THN_DENCL
KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH-
>THNX_TEST
KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS-YPNK
>1crnA
TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN-
```

clustal:

Petitioner Merck, Ex. 1066, p. 32

```
CLUSTAL W (1.82) multiple sequence alignment

THN_DENCL       KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH- 46
THNX_TEST       KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS-YPNK 46
1crnA           TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN- 46
                .:***  ..*  :  **: * .. :**  :** **..: **  *
```

msf:

```
 !!AA_MULTIPLE_ALIGNMENT 1.0

  thn_dencl.msf MSF:  47 Type: P 08/08/05 CompCheck:  427 ..

  Name: THN_DENCL  Len: 47  Check: 8212 Weight: 1.00
  Name: THNX_TEST  Len: 47  Check: 5295 Weight: 1.00
  Name: 1crnA      Len: 47  Check: 6920 Weight: 1.00

//

           1                                          47
THN_DENCL  KSCCPTTAARNQYNICRLPGTPRPVCAALSGCKIISGTGCPPGYRH~
THNX_TEST  KSCCPDTTGRDIYNTCRFGGGSRQVCARISGCKIISASTCPS.YPNK
1crnA      TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN~
```

Project Mode

In difficult modeling situations, where the correct alignment between target and template cannot be clearly determined by sequence based methods, visual inspection and manual manipulation of the alignment can significantly help improving the quality of the resulting model.

Project files contain the superposed template structures, and the alignment between the target and template. Project files can be generated inside the program DeepView (Swiss-PdbViewer *Guex et al.*), by the workspace template selection tools, and are also the default output format of the modeling pipeline. This allows analyzing and iteratively improving the the models generated by the "Automated mode" and "Alignment mode" modeling approaches.

The program DeepView can be downloaded freely from the tools section or from the ExPASy web site .

DeepView

The program DeepView (Swiss-PdbViewer, *Guex et al.*) can be used to generate, display, analyze and manipulate modeling project files for the SWISS-MODEL workspace.

Project files contain the superposed template structures, and the alignment between the target and template. The user has therfor full control over essential modelling parameters, i.e. the choice of template structures, the correct alignment of residues, and the placement of insertions and deletions in the context of the three-dimensional structure.

Project files can be generated inside DeepView, by the workspace template identification tools, and are also the default output format of the modeling pipeline. This allows analyzing and iteratively improving the output of the different modeling tools.

DeepView allows to visualize the model and the templates, and to analyse certain structural features e.g. Ramachandran plots or electrostatic properties. Moreover, it allows adjusting manually the placement of insertions and deletions in the alignment on which the initial modelling process was based on. The project with the modified alignment can then be re-submitted to the SWISS-MODEL workspace for model building.

DeepView can be downloaded at: http://www.expasy.org/spdbv/

DeepView does not require administrator privileges for installation. E.g. under MS windows, simply uncompress the distributed archive at any location you like (e.g. c:\spdbv or on your desktop) and start working by starting the spdbv.exe application.

Input target sequence and UniProt AC code

The amino acid sequence of a protein to be modeled or analyzed can be submitted in FASTA or raw format. If the protein sequence is deposited in the UniProt (*Bairoch et al.*)knowledgebase, the AC (ACcession number) for the entry can be also specified.

Examples:

- raw format: the amino acids sequence of the protein in plain-text:

Petitioner Merck, Ex. 1066, p. 33

```
MVEIVYWSGTGNTEAMANEIEAAVKAAGADVESVRFEDTNVDDVASKDVILLGCPAMGSE
ELEDSVVEPFFTDLAPKLKGKKVGLFGSYGWGSGEWMDAWKQRTEDTGATVIGTAIVNEM
PDNAPECKELGEAAAKA
```

- FASTA format consists of a single-line description, followed by lines of sequence data. The first character of the description line is a greater-than (">") symbol:

```
>sp|P00321|FLAV_MEGEL Flavodoxin - Megasphaera elsdenii.
MVEIVYWSGTGNTEAMANEIEAAVKAAGADVESVRFEDTNVDDVASKDVILLGCPAMGSE
ELEDSVVEPFFTDLAPKLKGKKVGLFGSYGWGSGEWMDAWKQRTEDTGATVIGTAIVNEM
PDNAPECKELGEAAAKA
```

- UniProt Accesion number: P00321

Display of modeling results

Coordinates of the model, the underlying alignment, log files, and quality evaluations can be accessed and downloaded via web-browser from the workspace.

Model Details

This section gives access to display the model and download its coordinates.

The model coordinates are available in two different formats:

- DeepView project files (recommended).
- PDB format

PDB formatted protein models can be displayed by any molecular visualization tool or browser-plugin. Here is a short list of freely available software:

- DeepView (MS Windows, Macintosh, Linux)
- DINO (Linux, IRIX, OSF,SUN)
- Rasmol (MS Windows, Mac, Unix)
- CHIME Plugin (requires registration)

If the model has been build using the *Automated Mode*, information about the template(s) used for modeling is provided with cross references to structural information databases via the link to the SWISS MODEL Template library.

Alignment Output

Displays the target template sequence alignment used in the modeling procedure and the assigned secondary structure.

Modeling Log

The modeling log gives a detailed description of the individual modeling steps. The models are built using the SWISS-MODEL server pipeline (*Schwede et al.*). The modelling log shows the individual steps during model building (*Guex et al.*), especially which parts of the model have been built ab initio (i.e. insertions / deletions).

Template Selection Log

The logfile provides information about the template selection step to search the SWISS-MODEL Template library for suitable templates.

Building of Homo-oligomeric assemblies

The quaternary structure annotation of the template is used to model the target sequence in its oligomeric form. The template complexes are derived by applying quaternary structure annotation given by the authors of the PDB entry (See PDB for more information). If such annotation is missing or ambiguous, the annotation of PISA [1] is used instead.

PISA estimates the stability of a complex by calculating a pseudo dissociation energy which includes interface stabilizing attributes (e.g. hydrogen bonds, salt bridges, disulfide bridges, hydrophobic interactions) but also entropic terms (e.g rotational and translational entropy). For a complete list of descriptors, please see [1].

A template is labeled as "homo-oligomeric" if the sequences of the "SEQRES" section (the sequences which were used to resolve the structure) are identical and "hetero-oligomeric" otherwise. The oligomeric state (e.g "MONOMERIC", "DIMERIC", etc.) is determined by counting all chains in the structure having more than 10 residues.

The model is built based on the quaternary form of the template structure, if conservation of the oligomeric state can be assumed with high confidence (i.e. <60% sequence identity between target and template sequences).

Otherwise a model is built in its monomeric form for the following reasons:

Petitioner Merck, Ex. 1066, p. 34

1. The template is annotated as monomer
2. The evolutionary relationship between the target and the template sequences is too low (Sequence identity < 60%).
3. The subunits in the complex are too different in structure or sequence, i.e. currently we restrict the computation to build homo-oligomeric assemblies.
4. The modeling routine of SWISS-MODEL fails to model the complex, e.g. too many loops to be reconstructed with de novo techniques.

[1] Krissinel E, Henrick K: Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007, 372(3):774-797.

Small molecules transfer

This method attempts to include a ligand to the modeled target sequence comparing the template's binding site residues with the corresponding ones in the target sequence.

To avoid including of ligands that are biologically irrelevant, only those within 3 Angstroem of any atoms of the template structure are evaluated. Additionally only ions that have at least 3 binding residues from a single chain or 2 binding residues from different chains within 3 Angstrom, are taken into account. An exception is made for ions that bind a cofactor, in this case they are joined to it and the resulting complex is treated as a single small molecule.

To find the residues that bind a ligand in both template and model, a structural alignment is performed with TMAlign [1] in order to get a overall superposition of the structures. TMAlign creates a pairwise alignment that is used to find the model's residues corresponding to the template's binding residues, which are the ones within 3 Angstrom around the ligand. After that, a superposition of only the binding sites is made to refine the structural alignment of the binding residues which will be evaluated; if it fails due to few residues being superposed, e.g. in case of ions, the procedure is repeated including template's residues that have the backbone within 14 Angstrom of the ligand, in order to be sure to include second shell residues.

The conservative approach used in this method defines that a ligand is included to the model when the following strict criteria are met: (1) the model's binding residues needs to be perfectly conserved, (2) the RMSD between the template's and model's binding residues is less than 2 Angstroem, and (3) there are no overlaps between the small molecule (which atom's positions are derived from the template structure) and the model or other ligands.

The small molecules taken into account until now in the pipeline are:
cations: CA, CD, CO, CU, CU2, FE, FE2, MG, MN, MO, NA, NI, ZN
cofactors: ADP, AMP, ATP, BTN, COA, BGC, GLC, GDP, GMP, GTP, GSH, FAD, FMN, HEM, HEA, HEB, NAD, NAP, NDP, NAI, PLP, SAM, THG, TPP, UDP, CDP, SF4, FES

The method is constantly improved, hence in the future more small molecules will be included and more physical-chemical properties, from the template and the model, will be used to decide in which case a certain ligand can be included into a model.

[1] Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on TM-score , Nucleic Acids Research, 2005 33: 2302-2309

Protein Structure & Model Assessment Tools

Evaluation of model quality is a crucial step in homology modeling. While the performance of the automated SWISS-MODEL (*Schwede et al.*) pipeline has been evaluated extensively by the EVA project (*Koh et al.*) and updates are benchmarked carefully, the quality of individual models can vary significantly.
Therefore, graphical plots of Anolea mean force potential (*Melo et al.*), GROMOS empirical force field energy (*van Gunsteren et al.*) and QMEAN (*Benkert et al.*) are provided to enable the user to estimate the local quality of the predicted structure. The stereo-chemistry of protein models and template structures can be analysed with Whatcheck (*Hooft et al.*) and Procheck (*Laskowski et al.*). In order to be able to rank alternative models of the same target protein, pseudo energies for the entire model as calculated by QMEAN (*Benkert et al.*) and DFIRE (*Zhou et al.*) are provided as well. To facilitate the description of template and model structures, DSSP (*Kabsch et al.*) and Promotif (*Hutchinson et al.*) can be invoked to classify structural features.

Anolea

The atomic empirical mean force potential ANOLEA (*Melo et al.*) is used to assess packing quality of the models. The program performs energy calculations on a protein chain, evaluating the "Non- Local Environment" (NLE) of each heavy atom in the molecule.

The y-axis of the plot represents the energy for each amino acid of the protein chain. Negative energy values (in green) represent favourable energy environment whereas positive values (in red) unfavourable energy environment for a given amino acid.

QMEAN is a composite scoring function for both the estimation of the global quality of the entire model as well as for the local per-residue analysis of different regions within a model.

QMEAN4 global score (SwissModel Workspace)
In the SwissModel Workspace the QMEAN4 score is used to evaluate the generated models. The global QMEAN4 scoring function (*Benkert et al. 2008*) is a linear combination of four structural descriptors using statistical potentials: The local geometry is analysed by a torsion angle potential over three consecutive amino acids. Two distance-dependent interaction potentials are used to assess long-range interactions: the first is a residue-level implementation based on C-beta atoms only and the second an all-atom potential which is able to capture more details of the model. A solvation potential investigates the burial status of the residues. The global QMEAN6 score uses two additional terms describing the agreement of the predicted (from sequence) and the calculated secondary structure and solvent accessibility of the model.

QMEAN4 is a reliability score for the whole model which can be used in order to compare and rank alternative models of the same target. The quality estimate ranges between 0 and 1 with higher values for better models. Additionally, the pseudo energies of the four contributing statistical potential terms are provided. The comparison of the differences of the terms among the models may help understanding the reason for the differences in the estimated model quality.

For the quality estimation of multiple models of the same protein, please visit the QMEAN server (*Benkert et al. 2009, NAR Web Server Issue*)

Petitioner Merck, Ex. 1066, p. 35

which allows to process sets of models (submitted as compressed archives) and pools the results: http://swissmodel.expasy.org/qmean

In addition to the raw scores, Z-scores of the QMEAN composite score as well as all terms are provided relating the quality estimates to scores obtained for high-resolution reference structures solved experimentally by X-ray crystallography (*Benkert et al. 2011*). The QMEAN Z-score represents an measure of the absolute quality of a model by providing an estimate of the 'degree of nativeness' of the structural features observed in a model and by describing the likelihood that a given model is of comparable quality to experimental structures. Models of low quality are expected to have strongly negative QMEAN Z-scores (i.e. the model's QMEAN score is several standard deviations lower than expected for experimental structures of similar size). The analysis of the Z-scores of individual terms may help identifying the geometrical features responsible for an observed negative QMEAN Z-score. A more detail explanation on how the Z-score are calculated can be found in the help of the QMEAN server.

QMEAN6 global score (Tool/Structure Assessment)

In "Tool/Structure Assessment" the QMEAN6 score is used. Compared to QMEAN4 used in SwissModel Workspace, QMEAN6 incorporates two additional terms witch investigate whether a model has the correct fold. These terms are useful, since, in contrast to the homology models from SwissModel, a model uploaded in Tool/Structure Assessment may be calculated by any method (e.g. physics-based or ab initio methods) with no guarantee that model was built based on a homologous template structure with the same fold.

The global QMEAN6 scoring function (*Benkert et al. 2008*) is a linear combination of six structural descriptors using statistical potentials: The local geometry is analysed by a torsion angle potential over three consecutive amino acids. Two distance-dependent interaction potentials are used to assess long-range interactions: the first is a residue-level implementation based on C-beta atoms only and the second an all-atom potential which is able to capture more details of the model. A solvation potential investigates the burial status of the residues. Two additional terms describing the agreement of the predicted (from sequence) and the calculated secondary structure and solvent accessibility of the model.

QMEAN6 is a reliability score for the whole model which can be used in order to compare and rank alternative models of the same target. The quality estimate ranges between 0 and 1 with higher values for better models. Additionally, the pseudo energies of the four contributing statistical potential terms are provided as well as the percentage agreement between predicted and measured features from the sequence and model, respectively. The comparison of the differences of the terms among the models may help understanding the reason for the differences in the estimated model quality.

For the quality estimation of multiple models of the same protein, please visit the QMEAN server (*Benkert et al. 2009, NAR Web Server Issue*) which allows to process sets of models (submitted as compressed archives) and pools the results: http://swissmodel.expasy.org/qmean.

In addition to the raw scores, Z-scores of the QMEAN composite score as well as all terms are provided relating the quality estimates to scores obtained for high-resolution reference structures solved experimentally by X-ray crystallography (*Benkert et al. 2011*). The QMEAN Z-score represents an measure of the absolute quality of a model by providing an estimate of the 'degree of nativeness' of the structural features observed in a model and by describing the likelihood that a given model is of comparable quality to experimental structures. Models of low quality are expected to have strongly negative QMEAN Z-scores (i.e. the model's QMEAN score is several standard deviations lower than expected for experimental structures of similar size). The analysis of the Z-scores of individual terms may help identifying the geometrical features responsible for an observed negative QMEAN Z-score. A more detail explanation on how the Z-score are calculated can be found in the help of the QMEAN server.

QMEAN: local score

The local version of the QMEAN scoring function (*Benkert et al. 2009*) consists of 8 terms (6 terms in the SwissModel Workspace). All terms are calculated over a sliding window of 9 residues and triangular smoothing is applied in order to put a stronger weight on the central residues of the window compared to the flanking ones. Adapted versions of the six terms (4 terms, respectively) used in the global version are combined with two additional features, namely, the average solvent accessibility (using triangular smoothing) and the fraction residues in the 9-residue window with no assigned secondary structure by DSSP. These two features take into account that, for example, solvent exposed loops are potentially less accurate than regions of regular secondary structure in the structural core of the protein.
The Residue Error Plot shows the local QMEAN score for each position in the model. The local score is an estimate of the expected structural inaccuracy at a given position with small values corresponding to regions in the model being potentially more reliable.

DFire

DFIRE (*Zhou et al.*) is an all-atom statistical potential based on a distance-scaled finite ideal-gas reference state. DFIRE is used to assess non-bonded atomic interactions in the protein model.

A pseudo energy for the entire model is provided which reflects the quality of the model and can be used for ranking alternative predictions of the same target. A lower energy indicates that a model is closer to the native conformation.

Gromos

GROMOS (*van Gunsteren et al.*) is a general-purpose molecular dynamics computer simulation package for the study of biomolecular systems and can be applied to the analysis of conformations obtained by experiment or by computer simulation.

The y-axis of the plot represents the energy for each amino acid of the protein chain. Negative energy values (in green) represent favourable energy environment whereas positive values (in red) unfavourable energy environment for a given amino acid.

What Check

What Check comprises several tools for protein structure verification (*Hooft et al.*).

Procheck

Petitioner Merck, Ex. 1066, p. 36

The PROCHECK suite of programs (*Laskowski et al.*) assess the "stereochemical quality" of a given protein structure. The aim of PROCHECK is to assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereochemical parameters derived from well-refined, high-resolution structures.

PROMOTIF

PROMOTIF (*Hutchinson et al.*) automatically identifies, classify and analyse a number of supersecondary structural motifs in proteins. Any resulting patterns will be useful in prediction of protein structure from amino acid sequence. Motifs analyzed include beta turns, gamma turns, Greek keys, beta hairpins and beta bulges. Data from PROMOTIF analyses are included in the PDBsum (*Laskowskiet et al.*) web site, which provides information derived from all currently available protein coordinate files.

DSSP

The DSSP (*Kabsch et al.*) program defines secondary structure, geometrical features and solvent exposure of proteins, given atomic coordinates in Protein Data Bank format. The program does NOT PREDICT protein structure.
The DSSP code

H = alpha helix
B = residue in isolated beta-bridge
E = extended strand, participates in beta ladder
G = 3-helix (3/10 helix)
I = 5 helix (pi helix)
T = hydrogen bonded turn
S = bend

SwissModel Template Library (ExPDB)

The template structure database used by SWISS-MODEL (SMTL or ExPDB library) is derived from the Protein Data Bank (*Westbrook et al.*). In order to allow sequence-based template searches, each PDB entry is split into individual chains. The separated template chains are annotated with information about experimental method, resolution (if applicable), ANOLEA mean force potential (*Melo et al.*), Gromos96 energy (*van Gunsteren et al.*) and PQS (*Henrick et al.*) quaternary state assignment to allow for rapid retrieval of the relevant structural information during template selection. Theoretical models, structures only consisting of C alpha atoms and irregularly formatted database entries are removed.
In order to speed up the sequence database search step of the template identification algorithms and to provide a clear and concise overview of the results, templates sharing 100% sequence identity are grouped into a SMTL100 library using the program CD-HIT, a fast clustering method for sequences at high identity thresholds (*Li et al.*). Clusters of sequences having 90%, 70% and 50% sequence identity are derived from the RCSB non-redundant PDB lists.

The ExPDB codes are constructed according to the following rule: PDBCODE+ChainID

Examples:
- *Light harvesting protein*: 1cpc contains two chains (with IDs A & B).
The corresponding ExPDB entries are respectively:

- Chain A: 1cpcA
- Chain B: 1cpcB

User specified template:

Depending on the planned model application, it can be necessary to select a different structural template than the one ranked first in the automated process. Typical examples are proteins in different conformational states, e.g. 1ake vs. 4ake. It is possible to specify the structure to be used as modelling template either by identifying an entry in the SWISS-MODEL template library by PDB-ID + ChainID e.g. "1ake" chain "A", or by uploading a file in PDB format (*) with coordinates of the template structure. Please make sure that this file contains only a single protein chain, and does not contain chemically modified amino acids, hereto atoms, ligands, etc.

(*) A simple PDB-like file containing the coordinates of the template structure. For more information about PDB file format please see link: http://www.wwpdb.org/docs.html

---

References:

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res 25(17): 3389-3402.

Andreeva, A., D. Howorth, et al. (2004). "SCOP database in 2004: refinements integrate structure and sequence family data." Nucleic Acids Res 32(Database issue): D226-9.

Bairoch, A., R. Apweiler, et al. (2005). "The Universal Protein Resource (UniProt)." Nucleic Acids Res 33 Database Issue: D154-159.

Eisenberg, D., R. Luthy, et al. (1997). "VERIFY3D: assessment of protein models with three-dimensional profiles." Methods Enzymol 277: 396-404.

Guex, N. and M. C. Peitsch (1997). "SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling." Electrophoresis 18(15): 2714-2723.

Hooft, R. W., G. Vriend, et al. (1996). "Errors in protein structures." Nature 381(6580): 272.

Petitioner Merck, Ex. 1066, p. 37

Hughey, R. and A. Krogh (1996). "Hidden Markov models for sequence analysis: extension and analysis of the basic method." Comput Appl Biosci 12(2): 95-107.

Hutchinson, E. G. and J. M. Thornton (1996). "PROMOTIF--a program to identify and analyze structural motifs in proteins." Protein Sci 5(2): 212-20.

Jones, D. T. (1999). "Protein secondary structure prediction based on position-specific scoring matrices." J Mol Biol 292(2): 195-202.
Jones, D. T. and J. J. Ward (2003). "Prediction of disordered regions in proteins from position specific score matrices." Proteins 53 Suppl 6: 573-578.
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
Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers 22: 2577-2637.

Söding J. (2005) "Protein homology detection by HMM-HMM comparison." Bioinformatics 21, 951-960. doi:10.1093/bioinformatics/bti125.

Koh, I. Y., V. A. Eyrich, et al. (2003). "EVA: Evaluation of protein structure prediction servers." Nucleic Acids Res 31(13): 3311-3315.

Laskowski R A, Chistyakov V V, Thornton J M (2005). PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nucleic Acids Res., 33, D266-D268.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993). 'PROCHECK: A program to check the stereochemical quality of protein structures' J. Appl. Cryst. 26: 283-291 (1993)

Li, W., L. Jaroszewski, et al. (2002). "Sequence clustering strategies improve remote homology recognitions while reducing search times." Protein Eng 15(8): 643-649.

Melo, F. and E. Feytmans (1998). "Assessing protein structures with a non-local atomic interaction energy." J Mol Biol 277(5): 1141-1152.

Zhou, H., and Zhou, Y. (2002). "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.
Benkert, P., Tosatto, S.C.E. and Schomburg, D. (2008). "QMEAN: A comprehensive scoring function for model quality assessment. " Proteins: Structure, Function, and Bioinformatics, 71(1):261-277.
Benkert P, Kuenzli M, Schwede T. (2009). "QMEAN Server for Protein Model Quality Estimation." Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W510-4.
Benkert, P., Schwede, T. and Tosatto, S.C.E. (2009). "QMEANclust: Estimation of protein model quality by combining a composite scoring function with structural density information." BMC Struct Biol. 2009 May 20;9:35.
Benkert P, Biasini M, Schwede T. (2011). "Toward the estimation of the absolute quality of individual protein structure models." Bioinformatics, 27(3):343-50.
Mulder, N. J., R. Apweiler, et al. (2005). "InterPro, progress and status in 2005." Nucleic Acids Res 33 Database Issue: D201-205.
Pearl, F., A. Todd, et al. (2005). "The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis." Nucleic Acids Res 33 Database Issue: D247-51.
Schwede, T., J. Kopp, et al. (2003). "SWISS-MODEL: An automated protein homology-modeling server." Nucleic Acids Res 31(13): 3381-3385.
van Gunsteren, W. F., S. R. Billeter, et al. (1996). Biomolecular Simulations: The GROMOS96 Manual and User Guide. Zürich, VdF Hochschulverlag ETHZ.
Henrick K, Thornton JM, PQS: "A protein quarternary file server." Trends Biochem. Sci. 1998;23:358-361.
Velankar, S., P. McNeil, et al. (2005). "E-MSD: an integrated data resource for bioinformatics." Nucleic Acids Res 33 Database Issue: D262-265.
Westbrook, J., Z. Feng, et al. (2003). "The Protein Data Bank and structural genomics." Nucleic Acids Res 31(1): 489-491.
Wheeler, D. L., T. Barrett, et al. (2005). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res 33 Database Issue: D39-45.
Zdobnov, E. M. and R. Apweiler (2001). "InterProScan--an integration platform for the signature-recognition methods in InterPro." Bioinformatics 17(9): 847-848.

SWISS-MODEL is developed by the Protein Structure Bioinformatics group at the SIB - Swiss Institute of Bioinformatics & the Biozentrum University of Basel. © 2011.

Petitioner Merck, Ex. 1066, p. 38

QMEAN Server for Model Quality Estimation

QMEAN Server - Quick Help

Introduction

Estimating the quality of protein structure models is a vital step in protein structure prediction. Often one ends up in having a set of alternative models (*e.g.* from different modeling servers or based on alternative template structures and alignments) from which the best candidate shall be selected. Or a singe model has been built from which the absolute quality needs to be predicted in order to have an idea about its suitability for subsequent experiments. The QMEAN server provides access to two scoring functions for the quality estimation of protein structure models which allow to rank a set of models and to identify potentially unreliable region within these. Both single models and set of models submitted as *tar.gz*-archives can be analysed. The user has the possibility to choose between the following two scoring functions:

- **QMEAN** [1,3] is a composite scoring function which is able to derive both global (*i.e.* for the entire structure) and local (*i.e.* per residue) error estimates on the basis of *one single model*.
  NEW Recently (manuscript in preparation), the QMEAN score has been extended to an absolute quality estimate (see section "Estimated absolute model quality" below).
- **QMEANclust** [3] derives the score for a model by analysing its structural difference to all other models in the *ensemble*. The basic idea behind it is, that structural features observed more frequently have a higher probability to be correct. The initial ranking obtained by QMEAN is thereby used to weight the contribution of each model in the calculation of the QMEANclust consensus score.

The accuracy of the QMEANclust quality estimation improves with the size, the diversity (models from different severs, models based on different templates etc.) and the quality (fraction of near-native structures) of the model ensemble to be analysed. In order to obtain meaningful results a minimum number of models should be provided (*e.g.* > 30 models).

Input format requirements

Either single models (PDB-format) or *tar.gz*-archives with multiple models of the same protein can be uploaded. Additionally, if more than one model is submitted, the full-length sequence of the protein has to be provided (as sequence string or in FASTA format). In the case of multiple models, the sequences of the models are mapped on the target sequence and the models are automatically renumbered if necessary. A flag can be set in order to penalise incomplete models. In this case, the model score is additionally multiplied by the fraction of modelled residues with respect to the input sequence thereby punishing short models.

Results page

As an demonstration, pre-calculated example results from the CASP7 blind test experiment are shown below and are accessible interactively on the following page.

Petitioner Merck, Ex. 1066, p. 39

Project:	61 CASP models from different servers (target T0308, first model)
Model quality estimation method used	QMEAN
Number of structures processed	57
Number of structures skipped	4
Penalize incomplete models	no

Additional downloads:

- **All results in a single tar.gz-archive**
- Table with QMEAN scores (.csv)
- Table with QMEAN Z-scores w.r.t reference PDBs (.csv) **NEW**
- QMEAN detail table (.csv): all energy terms contributing to QMEAN **NEW**
- QMEAN output in CASP format

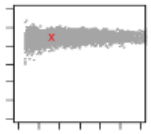
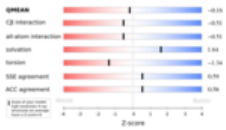
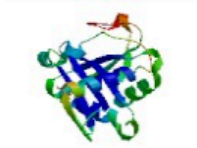
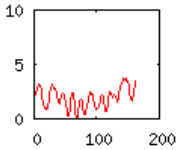
The results page begins with an *summary* on the input data and provides compact tables concerning the model ensemble. *E.g.* a table is provided which lists for each model the values of the 6 scoring function terms contributing to QMEAN. A short description of all terms can be found in the table below:

Scoring Function	Description
torsion	Extended torsion potential over 3 consecutive residues. Bin sizes: 45 degree for the center residue, 90 degree for the 2 adjacent residues.
pairwise	Residue-level, secondary structure specific interaction potential using Cβ atoms as interaction centres. Range 3...25 Å, step size: 1 Å
solvation	Potential reflecting the propensity of a certain amino acid for a certain degree of solvent exposure approximated by the number of Cβ atoms within a sphere of 9 Å around the centre Cβ.
all_atom	All-atom, secondary structure specific interaction potential using all 167 atom types. Range 3...20 Å, step size: 0.5 Å
SSE_agree	Agreement between the predicted secondary structure of the target sequence (using PSIPRED) and the calculated secondary structure of the model (using DSSP).
ACC_agree	Agreement between the predicted relative solvent accessibility using ACCpro (buried/exposed) and the relative solvent accessibility derived from DSSP (>25% accessibility => exposed)
QMEAN	The original QMEAN score as published in Benkert *et al.* 2008. It consists of a linear combination of all the six terms described above. The original QMEAN score has been replaced by the QMEANnorm score in all calculations and is only used in the detail table in order to be able to compare the scores and differentiate between them.
QMEANnorm **NEW**	Composite score in analogy to QMEAN but based on normalized statistical potential terms. The normalisation reduces the dependence of the quality score on the size of the model (such that larger proteins do not automatically get assigned higher absolute scores). The QMEANnorm score builds the basis of all calculations described below (model ranking, Z-scores). Whenever the term QMEAN is used on the webpage, it refers to the QMEANnorm score described below *(manuscript in preparation)*.

The inspection of the differences of the terms between the models may help understanding which terms contributed most to the low quality estimate of a certain model. For the four statistical potential terms, lower pseudo energies mean higher reliability. The QMEAN score as well as the two agreement terms range from 0 to 1 with higher values for more reliable candidates. In the case of QMEANclust, "local conformational diversity" plots showing the median QMEANclust score per position are provided which help to analyse the diversity within the ensemble of models.

Since the QMEAN score is protein-size dependent (*i.e.* larger proteins tend to have higher scores), the **QMEANnorm** score has been introduced *(manuscript in preparation)*.. In QMEANnorm, the four statistical potential terms are normalized: the interaction energy is divided by the total number of interactions and the other two (single-body) terms are normalized by the protein size. The QMEAN detail table provided in the summary section contains both the original statistical potentials terms and their normalized counterparts together with the QMEAN and the QMEANnorm scores. Unless not specified otherwise, everywhere the term QMEAN is used on the webpage, it refers to the QMEANnorm score described above which is the primary quality score used to calculate the Z-score and rank the models.

Petitioner Merck, Ex. 1066, p. 40

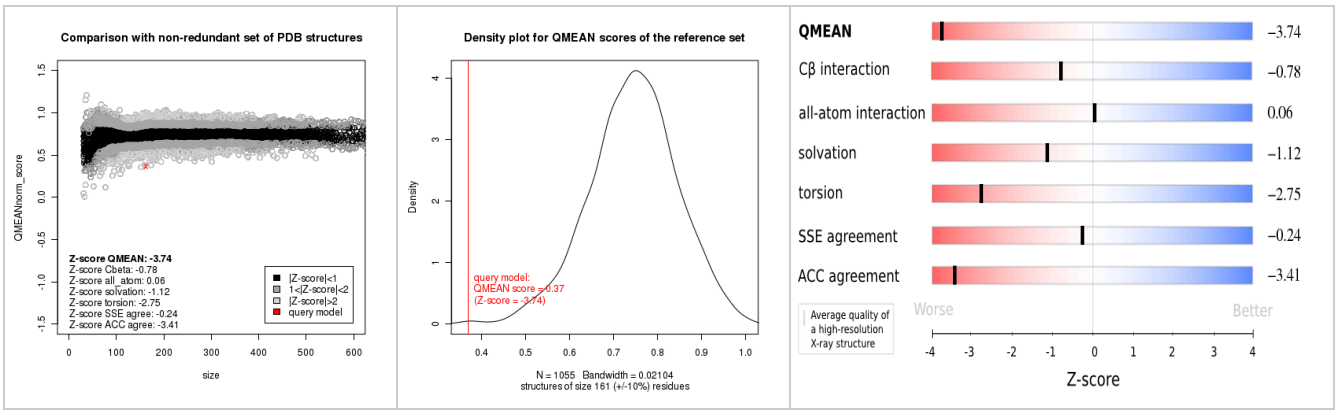
Global scores				Local scores	
Model name	QMEAN score	Estimated absolute model quality NEW	Z-scores of QMEAN terms NEW	Residue error <1Å >3.5Å	Residue error plot
BAKER-ROBETTA_TS1_T0400	0.689	 Z-score=-0.15 [plot 1] [plot 2]	 [png]	 [jpg] [pdb] Jmol	 [png] [ps] [table]

For each model the following data and plots are provided in separate columns of the output table (depending on the quality estimation method, the ranking of the models is based on the QMEAN or QMEANclust score):

- **Model name**: filename as given in the *tar.gz*-archive
- **QMEAN score / QMEANclust score**: global score of the whole model reflecting the predicted model reliability ranging from 0 to 1.
- **Estimated absolute model quality**: The QMEAN score of the query model is related to the scores of a non-redundant set of high-resolution X-rays structures of similar size and a Z-score is calculated (more details below).
- **Residue error**: The estimated residue error is visualised using a colour gradient from blue (more reliable regions) to red (potentially unreliable regions, estimated error above 3.5 Å). The per residue error is written in the B-factor column (*pdb*-file or coloured model as *jpeg* for download). The molecular graphics viewer Jmol *(http://www.jmol.org/)* can be directly used on the website to interactively inspect the problematic regions in the colour-coded structure.
- **Residue error plot**: model energy profile with estimated residue errors along the sequence (postscript and png-file for download)
- **Energy profiles**: The local model quality data are also provided as tables in tab-separated format. The first table contains for each residue the values of the terms building the QMEAN scoring function. The second table provides the QMEAN/QMEANclust score per residue.

The range of local error estimates varies considerably between the two local versions of QMEAN and QMEANclust. QMEANlocal is, as a consequence of the statistical potential terms used, unable to discriminate between serious and very serious deviations (*e.g* between 5 Å and 15 Å). QMEANclust on the other hand can, depending on the quality and size of the ensemble, provides error estimates for even large errors.

NEW Estimated absolute model quality



The **QMEAN Z-score** [2] provides an estimate of the absolute quality of a model by relating it to reference structures solved by X-ray crystallography. The QMEAN Z-score is an estimate of the "degree of nativeness" of the structural features observed in a model by describing the likelihood that a model is of comparable quality to high-resolution experimental structures.
The three plots available for download visualize the quality of a given model with respect to these reference structures. The reference structures are a non-redundant subset of the PDB sharing less than 30% pairwise sequence identity among each other and are solved at a resolution below than 2 Å.

Plot 1:

Petitioner Merck, Ex. 1066, p. 41

The area built by the circles colored in different shades of grey in the plot on the left hand side represent the QMEAN scores of the reference structures from the PDB. The model's QMEAN score is compared to the scores obtain for experimental structures of similar size (model size +/- 10%) and a Z-score is calculated. A Z-score (or standard score) is a score which is normalised to mean 0 and standard deviation 1. Thus the QMEAN Z-score directly indicates how many standard deviations the model's QMEAN score differs from expected values for experimental structures. In analogy, Z-scores are calculated for all four statistical potential terms as well as the agreement terms being part of the QMEAN score (see also Plot 3).

Plot 2:

The plot in the middle shows the density plot (based on the QMEAN score) of all reference models used in the Z-score calculation. The location of the query model w.r.t. the background distribution is marked in red. This plot basically is a "projection" of the first plot for the given protein size. The number of reference models used in the calculation is shown at the bottom of the plot.

Plot 3:

The analysis of these Z-scores of the individual terms can help identifying the geometrical features responsible for an observed large negative QMEAN Z-score. Models of low quality are expected to have strongly negative Z-scores for QMEAN but also for most of the contributing terms. Large negative values correspond to red regions in the color gradient. "Good structures" are expected to have all sliders in the light red to blue region.

Note:

The quality estimates for **membrane proteins** need to be treated with caution: Membrane proteins may receive very low Z-scores since their physico-chemical properties differ considerably from those of soluble proteins. A QMEAN version with separate potentials optimised for membrane proteins is under development.

References

Reference for the QMEAN scoring function:
[1] Benkert, P., Tosatto, S.C.E. and Schomburg, D. (2008). "QMEAN: A comprehensive scoring function for model quality assessment." Proteins: Structure, Function, and Bioinformatics, 71(1):261-277. PubMed

Reference for the QMEAN Z-scores:
[2] Benkert, P., Biasini, M. and Schwede, T. (2011). "Toward the estimation of the absolute quality of individual protein structure models." Bioinformatics (2010). doi: 10.1093/bioinformatics/btq662 PubMed

Reference for the QMEANclust scoring function:
[3] Benkert, P., Schwede, T. and Tosatto, S.C.E. (2009). "QMEANclust: Estimation of protein model quality by # combining a composite scoring function with structural density information." BMC Struct Biol. 2009 May 20;9:35. PubMed

Reference for the QMEAN server:
[4] Benkert P, Künzli M, Schwede T. (2009). "QMEAN Server for Protein Model Quality Estimation." Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W510-4. PubMed

QMEAN is developed by the Protein Structure Bioinformatics group at the SIB - Swiss Institute of Bioinformatics & the Biozentrum University of Basel. © 2010.

Petitioner Merck, Ex. 1066, p. 42

The Wayback Machine - https://web.archive.org/web/20110519144341/http://swissmodel.expasy.org/repository/?pid=smd08

SWISS-MODEL

Visualisation

The following tools allow visualization of macromolecular structures in PDB format (i.e. experimental structures and models). SWISS-MODEL results are best viewed and manipulated with the program DeepView (Swiss-PdbViewer), which also allows to modify the target template alignment for the SWISS-MODEL "project mode".

- **DeepView** The Swiss-PdbViewer
- **DINO** Protein Structure Visualization for Linux
- **Molscript** a program for displaying molecular 3D structures such as proteins.
- **POV-RAY** Open Source Raytracing software
- **JMOL** an open-source Java viewer for chemical structures in 3D
- **PyMOL** is a Python based open-source viewer for visualization of macromolecular structures
- **Chimera** extensible program for interactive visualization and analysis of molecular structures

Petitioner Merck, Ex. 1066, p. 43

SWISS-MODEL is developed by the Protein Structure Bioinformatics group at the SIB - Swiss Institute of Bioinformatics & the Biozentrum University of Basel. © 2011.