Bioinformatics Advance Access published November 12, 2005

BIOINFORMATICS

The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling

Konstantin Arnold[1,2], Lorenza Bordoli[1,2], Jürgen Kopp[1,2] and Torsten Schwede[1,2]

[1]Biozentrum Basel, University Basel, Switzerland and [2] Swiss Institute of Bioinformatics, Basel, Switzerland

ABSTRACT

Motivation: Homology models of proteins are of great interest for planning and analyzing biological experiments when no experimental three-dimensional structures are available. Building homology models requires specialized programs and up-to-date sequence and structural databases. Integrating all required tools, programs and databases into a single web-based workspace facilitates access to homology modelling from a computer with web connection without the need of downloading and installing large program packages and databases.

Results: 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. A personal working environment is provided for each user where several modelling projects can be carried out in parallel. Protein sequence and structure databases necessary for modelling are accessible from the workspace and are updated in regular intervals. Tools for template selection, model building, and structure quality evaluation can be invoked from within the workspace. Workflow and usage of the workspace are illustrated by modelling human Cyclin A1 and human Transmembrane Protease 3.

Availability: The SWISS-MODEL Workspace can be accessed freely at http://swissmodel.expasy.org/workspace/.

Contact: Torsten.Schwede@unibas.ch

1 INTRODUCTION

Three-dimensional protein structures are of great interest for the rational design of many different types of biological experiments, such as site-directed mutagenesis or structure-based discovery of specific inhibitors. However, the number of structurally characterized proteins is small compared to the number of known protein sequences: as of November 2005, more than 33'000 experimentally determined protein structures were deposited in the Protein Data Bank (Westbrook, Feng et al. 2003), while the UniProt protein knowledge database (Bairoch, Apweiler et al. 2005) held more than 2.3 million sequences. Various computational methods for modelling three-dimensional structures of proteins have been developed to overcome this limitation. Since the number of possible folds in nature appears to be limited (Chothia 1992) and the three-dimensional structure of proteins is better conserved than their sequences, it is often possible to identify a homologous protein with a known structure (template) for a given protein sequence (target). In these cases, homology modelling has proven to be the method of choice to generate a reliable three-dimensional model of a protein from its amino acid sequence as impressively shown in several meetings of the bi-annual CASP experiment (Tramontano and Morea 2003; Moult 2005). Homology modelling is routinely used in many applications, such as virtual screening, or rationalizing the effects of sequence variations (Marti-Renom, Stuart et al. 2000; Kopp and Schwede 2004; Hillisch, Pineda et al. 2004). Building a homology model comprises four main steps: 1) identification of structural template(s), 2) alignment of target sequence and template structure(s), 3) model building, and 4) model quality evaluation. These steps can be repeated until a satisfying modelling result is achieved. Each of the four steps requires specialized software as well as access to up-to-date protein sequence and structure databases.

Selection of the most suitable template structure and the alignment between target and template are still the predominant sources of errors in comparative models (Tramontano and Morea, 2003). In our own experience and that of others (Bates et al., 2001), manual intervention at different steps of model building can significantly improve the accuracy of the results. We have developed the SWISS-MODEL Workspace to address this aspect by providing a range of tools that allow the user to validate and modify the different modelling steps manually. The SWISS-MODEL Workspace integrates programs and databases required for homology modelling in an easy-to-use web-based modelling workbench. It allows the user to construct comparative protein models from a computer with web connection without the need of downloading and installing large program packages and databases.

Depending on the difficulty of the individual modelling task, the workspace assists the user in building and evaluating protein homology models at different levels of complexity. For models that can be built based on sufficiently similar templates, we provide a highly automated modelling procedure with a minimum of user intervention. On the other hand, for more difficult modelling scenarios with less target-template sequence similarity, the user is given control over individual steps of model building: functional domains in multi-domain proteins can be detected, secondary structure and disordered regions can be predicted for the target sequence, suitable template structures identified by searching the SWISS-MODEL Template Library (SMTL), and target-template

*To whom correspondence should be addressed.

alignments can be manually adjusted. As quality evaluation is indispensable for a predictive method like homology modelling, every model is accompanied by several quality checks. All relevant data of a modelling project is presented in graphical synopsis.

The accuracy, stability and reliability of the automated SWISS-MODEL server pipeline (Peitsch 1995; Guex and Peitsch 1997; Schwede, Kopp et al. 2003) has been continuously validated by the EVA-CM evaluation project (Koh et al., 2003). The Workspace extends the web-based SWISS-MODEL system by assisting manual user intervention in model building and evaluation. Here, we will illustrate the application of SWISS-MODEL Workspace in two modelling projects of different complexity: human Cyclin A1 and Transmembrane Protease 3.

2 SYSTEMS AND METHODS

2.1 Modelling modes in SWISS-MODEL Workspace

Depending on the difficulty of the modelling task, three different types of modelling modes are provided which differ in the amount of user intervention: automated mode, alignment mode, and project mode. Modelling requests are directly computed by the SWISS-MODEL server homology modelling pipeline (Schwede, Kopp et al. 2003). The “automated mode” has been developed 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 identical residues (Rost 1999). “Automated mode” submissions require only the amino acid 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. The automated template selection will favour high-resolution template structures with reasonable stereochemical properties as assessed by ANOLEA mean force potential (Melo and Feytmans 1998) and Gromos96 force field energy (van Gunsteren, Billeter et al. 1996).

Multiple sequence alignments are a common tool in many molecular biology projects, and are often the result of extensive theoretical and experimental exploration of a certain family of proteins. If a three-dimensional structure for at least one of the members is known, a given alignment can be used as starting point for comparative modelling using the “alignment mode”. In order to facilitate the use of alignments, the submission is implemented as a three step procedure: The input alignment, which can be provided in several different formats (FASTA, MSF, ClustalW, PFAM, SELEX), is converted into a standard ClustalW format in the first step. The user indicates which sequence in the alignment corresponds to the target protein, and which corresponds to a protein with an experimentally determined structure deposited in the PDB database (Westbrook, Feng et al. 2003). Commonly, protein sequences used for multiple sequence alignments are not identical to their corresponding PDB entries, e.g. due to missing atom coordinates, mutations, or sequence tags introduced for easier purification and crystallization. Therefore, in the subsequent step the submitted alignment is matched against the sequence of the template structure coordinates extracted from the SWISS-MODEL Template Library. Since most modelling programs depend on a perfect match between the template sequence in the input alignment and the residues present in the coordinate file, this step relieves the user from the obligation to deliver "perfect" alignments to PDB entries. Instead one can actually work with sequences directly retrieved from sequence databases. The server pipeline will build the model based on this alignment. During the modelling process, implemented as rigid fragment assembly in the SWISS-MODEL pipeline, the modelling engine might introduce minor heuristic modifications to improve the placement of insertions and deletions based on the structural context (Schwede, Kopp et al. 2003).

In difficult modelling projects, 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 improve the quality of the resulting model (Bates, Kelley et al. 2001). The program DeepView - Swiss-PdbViewer - (Guex and Peitsch 1997) can be used to generate, display, analyze and manipulate modelling project files for SWISS-MODEL workspace. Project files contain the superposed template structures and the alignment between the target and template. In this mode the user has 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 also be generated by the workspace template selection tools and are the default output format of the modelling pipeline. This allows analyzing and iteratively improving the output of “automated mode” and “alignment mode” modelling projects. The program DeepView can be downloaded freely from the ExPASy web site (http://www.expasy.org/spdbv/).

2.2 SWISS-MODEL Template Library

The template structure database used by SWISS-MODEL (SMTL) is derived from the Protein Data Bank (Westbrook, Feng et al. 2003). 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, Gromos96 energy and PQS (Henrick and Thornton 1998) quaternary state assignment to allow for rapid retrieval of the relevant structural information during template selection. Theoretical models, structures only consisting of C_α atoms and incorrectly 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, Jaroszewski et al. 2001; Li, Jaroszewski et al. 2002). Clusters of sequences having 90%, 70% and 50% sequence identity are derived from the RCSB non-redundant PDB lists. 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, Feng et al. 2003), SCOP (Andreeva, Howorth et al. 2004), CATH (Pearl, Todd et al. 2005), PDBsum (Laskowski, Chistyakov et al. 2005), and MSD (Velankar, McNeil et al. 2005).

2.3 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 (Andreeva, Howorth et al. 2004; Bateman, Coin et al. 2004; Pearl, Todd et al. 2005). The member databases of InterPro (Mulder, Apweiler et al. 2005) allow for both the identification of protein domains and the assignment of protein function. Using IprScan (Zdobnov and Apweiler 2001), a PERL-based InterProScan implementation, protein domains and functional sites can be assigned to regions of a target sequence. Splitting multi-domain proteins into separated domains often improves the sensitivity and performance of profile-based template search methods (see below). Secondary structure prediction methods are provided, which 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. The secondary structure prediction program PsiPred (Jones 1999) is accessible from the tools section of the Workspace.

Most soluble proteins adopt well-defined three-dimensional structures. However, some proteins have regions that are natively disordered, unstructured or have flexible regions without permanent regular secondary structure. It has been suggested that disordered regions may possess biological functions, and could be involved in signalling and regulation processes (Dunker and Obradovic 2001; Iakoucheva, Brown et al. 2002). The protein disorder prediction program Disopred (Jones and Ward 2003) estimates the propensity of protein sequences to be disordered. The result of secondary structure, disorder and domain boundary prediction (see figure 1) can aid the selection of modelling templates for specific regions of the target protein.

2.4 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 a so far unknown fold. Consequently, the SWISS-MODEL Workspace provides access to a set of increasingly complex and computationally demanding methods to search for templates.

Blast. Templates which are close homologues of the target can be identified using a gapped BLAST (Altschul, Madden et al. 1997) query against the SWISS-MODEL template library extracted from PDB. When no suitable templates are identified, or only parts of the target sequence are covered, two additional approaches for more sensitive detection of distant relationships among protein families are provided: an iterative profile Blast search and a Hidden Markov Model based template search.

Iterative Profile Blast. In the iterative profile Blast approach (Altschul, Madden et al. 1997) a profile for the query sequence is build from homologues sequences found by iterative searches of the NR database (Wheeler, Barrett et al. 2005). Subsequently, this profile is used to search for homologous structures in the template library. This method has been initially introduced as PDB-Blast by Godzik and coworkers.

HMM based template library search. A library of hidden markov models for a non-redundant set of the sequences in the template library was generated. Each model of the library was created from a multiple sequence alignment generated by an iterative search of the NR databases using SAM-T2K (Hughey and Krogh 1996). Model building, calibration and library searches were performed using the SAM (v 3.4) software package (Karplus, Barrett et al. 1998). Although computationally intensive, model calibration against a large, non-redundant database was calculated in order to obtain more accurate E-values during the scoring step. To select suitable template structures, the target sequence is scored against the template HMM library for statistically significant matches.

2.5 Continuous evaluation of modelling accuracy

Evaluation of model quality is a crucial step in homology modelling. In our view, benchmarking the individual steps and components of the comparative modelling workflow separately is not sufficient to assess the overall performance of a modelling pipeline, because the observed differences may not reflect critical steps in practical application and are therefore often inconclusive in judging the accuracy of an entire modelling workflow. The reliability of different protein modelling methods can be assessed by evaluating the results of blind predictions after the corresponding protein structures have been determined experimentally. During the biannual "Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction" CASP (Moult, 2005), new algorithmic developments are evaluated based on a number of test cases, e.g. in the 2004 CASP6 experiment 43 homology modelling targets were assessed (Tress *et al.*, 2005). The continuous and automated assessment of modelling servers by the LiveBench (Rychlewski and Fischer, 2005) or EVA (Koh et al., 2003) projects is based on a large number of blind predictions. SWISS-MODEL was the first comparative modelling server to join the EVA project in May 2000, and has been continuously monitored since then. EVA-CM has performed a comparative evaluation of the following servers: 3D-Jigsaw (Bates et al., 2001), ESyPred3D (Lambert et al., 2002), CPHModels (Lund et al., 1997), SDSC1 (Shindyalov and Bourne, 2000), and SWISS-MODEL (Guex and Peitsch, 1997; Schwede et al., 2003). The evaluation has been based on 261 weekly releases of the PDB database consisting of 48'098 protein models for 19'698 protein target chains. Figure 2a (Suppl. material) gives an estimation of the overall accuracy of the different modelling servers displaying the C_α atom RMSD after global superposition of the model and the experimental target structures vs. % of sequence identity between target and best template. As expected, model RMSD is increasing with decreasing alignment accuracy as defined by the percentage of equivalent C_α positions (within 3.5 Angstroms) between the optimally superimposed target and model structures (Figure 2b, suppl. material). For target-template pairs sharing sequence identities of more than 40 %, only small differences between the prediction servers are observed.

The cumulative evaluation of global C_α atom RMSD over all models in the test set shows that the models build by SWISS-MODEL exhibit on average the lowest overall deviation (2.0 Å vs. 2.7 Å for 3DJIGSAW) from the experimental control structures (Figure 2c, suppl. material). However, this is linked to a lower coverage of SWISS-MODEL for low homology templates. In the EVA-CM comparison, the models build by SWISS-MODEL ap-

pear to be on average more accurate, but shorter than the models of the other servers in the evaluation.

In general, the modelling accuracy for different target proteins is much more variable than the differences observed between different modelling methods for the same target. The EVA-CM database is sufficiently large to allow the prediction of the expected accuracy of the participating servers for a given protein sequence based on previous data of similar modelling scenarios.

2.6 Tools for protein structure and model assessment

The quality of individual models can vary significantly from the average accuracy expected for a given target-template similarity or modelling method. Therefore, individual assessment of each model is essential. SWISS-MODEL Workspace provides graphical plots of Anolea mean force potential (Melo and Feytmans, 1998), GROMOS empirical force field energy (van Gunsteren et al., 1996), Verify3D profile evaluation (Eisenberg *et al.*, 1997), Whatcheck (Hooft *et al.*, 1996) and Procheck (Laskowski *et al.*, 1993) reports are generated to enable the user to estimate the quality of protein models and template structures. To facilitate the description of template and model structures, DSSP (Kabsch and Sander, 1983) and Promotif (Hutchinson and Thornton, 1996) can be invoked to classify structural features.

2.7 Implementation

We have implemented the SWISS-MODEL Workspace using standard web browsers as graphical user interface to the dynamic server front-end based on Apache/PHP. Protein sequence and structure databases, modelling software, and user data are located on a back-end system. Front-end server and the compute back-end communicate via XML-RPC calls, separating the graphical display from the computationally demanding tasks. Batch jobs are executed on a high-performance Linux cluster via Sun GRID engine to ensure load balancing and optimal response times. Databases at the back-end are automatically updated in regular intervals. The SWISS-MODEL Workspace provides a personal working area for each registered user, which is not visible for other users. Email notification is sent to indicate when the calculation of a work unit has been completed. Users can access the workspace system anonymously without providing an email address; however, it is then necessary to bookmark the individual work units in the web browser to be able to retrieve the results once the browser has been closed. The SWISS-MODEL workspace is accessible at http://swissmodel.expasy.org/workspace/.

3 RESULTS AND DISCUSSION

In the following chapter, we describe how the different modeling modes and tools available from the SWISS-MODEL Workspace were applied to identify suitable templates and to build homology models for two biologically relevant human proteins: Cyclin A1 (CCNA1) and Transmembrane Protease 3 (TMPRSS3).

3.1 Modelling of human Cyclin A1

Cyclins are eukaryotic proteins that play an active role in controlling nuclear cell division cycles, and regulate cyclin dependent kinases (CDKs). Inactive CDK apoenzymes are partially activated by complex formation with regulatory Cyclin subunits and the complexes are further activated by phosphorylation of a Thr residue in CDKs. There are four classes of cyclins: A, B, D, E, which bind to CDKs at different stages of the cell cycle. The mammalian A-type cyclin family consists of two members, Cyclin A1 and Cyclin A2. While Cyclin A2 is widely expressed in different tissues (Liu, Matzuk et al. 1998), Cyclin A1 is limited to male germ cells. Cyclin A1 is essential for spermatocyte passage into the first meiotic division in male mice (Liu, Matzuk et al. 1998). Therefore its role in human male infertility is explored and it is investigated whether it could be a novel target for new approaches for male contraception (Wolgemuth, Lele et al. 2004).

To date no experimentally determined three-dimensional structure of Cyclin A1 is available. Blast searches to identify suitable templates for the modeling of Cyclin A1 lead to several highly significant matches. These matches span the second half of the protein (starting at about residue 210), which correspond to the Cyclin domain, as assigned by scanning the InterPro database of protein families and domains. All detected templates belong to the Cyclin family; in particular the Cyclin A2 proteins share the highest sequence similarity with Cyclin A1. For Cyclin A2 several experimental structures are known, which contain two subdomains of similar all-alpha fold forming the Cyclin domain of the protein.

The Blast alignment between the sequences of the Cyclin domain of Cyclin A1 and Cyclin A2 is unambiguous: the two sequences share about 60% identity and contain only one gap. Based on these observations, the cyclin A1 regulatory domain can be easily modeled using the automated mode of the SWISS-MODEL pipeline. The modelling pipeline selects the structure of the human Cyclin A2 in complex with CDK2 and an inhibitor (PDB: 1h1rB (Davies, Bentley et al. 2002)) as a template. Although other structures of the same protein have been deposited in the PDB, 1h1rB is selected by the modelling pipeline because it has the highest resolution among the suitable templates. Alternatively, other Cyclin A2 apo-protein structures or complexes with other molecules could be used to model the Cyclin A1 C-terminal domain. Exhaustive information about the individual templates is available directly from the template selection output as link to the SWISS-MODEL template library and external resources: MSD (Velankar, McNeil et al. 2005), PDB (Westbrook, Feng et al. 2003), PDBsum (Laskowski, Chistyakov et al. 2005), SCOP (Andreeva, Howorth et al. 2004) and CATH (Pearl, Todd et al. 2005). For instance, in order to model the Cyclin-CDK activated complex bound to ATP, the bovin template 1finB (Jeffrey, Russo et al. 1995) could be used instead of 1h1rB. In the automated mode of the modelling pipeline, it is sufficient to specify the SMTL code for the template to be used.

As expected from the high target-template similarity the quality of both models (based on structures 1h1rB and 1finB) is very good. Overall ANOLEA, GROMOS and Verify3D quality assessments do not detect major problems in the models. Visual inspection of the model and the template structure using DeepView shows good sequence conservation at the interface between CDK2 and the Cyclin molecules.

3.2 Modeling of TMPRSS3

While the automated approach yields satisfying results for closely related proteins, modelling of proteins based on templates with remote homology requires more user intervention as presented in the case of TMRSS3. Human Transmembrane Protease 3

(TMPRSS3) belongs to a family of transmembrane serine proteases expressed in several tissues. Mutations in the protein have been reported to cause neurosensory deafness (Masmoudi, Antonarakis et al. 2001; Scott, Kudoh et al. 2001; Wattenhofer, Di Iorio et al. 2002; Lee, Park et al. 2003). In silico predictions indicate that the protein contains four domains: a transmembrane (TM) domain, a low-density lipoprotein receptor A (LDLRA), a scavenger receptor cystein-rich (SRCR) domain and a protease domain. Pathogenic mutations have been described in all three non-membrane domains. Currently, no experimentally determined protein structure of human TMPRSS3 is available. Therefore, homology models of the protein are valuable resources to rationalize the functional impact of these mutations.

The boundaries of the individual domains of TMPRSS3 were identified using IprScan. The sensitivity and performance of template detection can often be improved when the template search is performed on individual domains rather than the whole target sequence. As previously reported (Masmoudi, Antonarakis et al. 2001; Scott, Kudoh et al. 2001; Wattenhofer, Di Iorio et al. 2002; Lee, Park et al. 2003), IprScan reports the occurrence of LDLRA, SRCR and protease domains in the TMPRSS3 protein. Blast, Iterative Profile Blast and HMM based template identification methods display several matches to structures corresponding to the protease domain and some templates spanning both SRCR and protease domain. For the LDLRA domain only Iterative Profile Blast and HMM based template identification methods are sensitive enough to detect potential structural templates with significant scores (Fig. 1). Each template of the hit list is linked to the SMTL and from there to other external resources, to allow for verification and a plausibility check. The biological annotations (e.g. domain classifications) of the different potential templates can be compared with the annotations of the target sequence to avoid the selection of non-homologous templates. This is especially relevant when trying to increase the coverage of the model by including low homology templates. The templates covering both SRCR and protease domain correspond to the structure of human Hepsin which belongs to the family of serine proteases. Crystal structures of a protein-substrate complex (PDB: 1z8g, (Herter, Piper et al. 2005)) or the apo-protein (PDB: 1P57, (Somoza, Ho et al. 2003)) could be used to build a model for the TMSSR3 protease and SRCR receptor domains. The alignment between target sequence and template 1z8g contains several gaps and therefore should be checked carefully. Alignments with possible template structures can be directly loaded into DeepView where the placement of insertion and deletions is examined in the structural context. Modified alignments can be saved as project files and submitted to the Project mode of the modeling workspace. The protein structure deposited as PDB entry 1z8g contains the cleaved activated form of the protease. To build a model of the activated form, the two fragments of the template were treated as individual chains within DeepView. Several alternative alignments were submitted and the quality of the resulting models was inspected based on the output of protein model assessment tools, which accompany the resulting models. Geometric correctness of the models e.g. Ramachandran plots were analyzed with Procheck (Laskowski, MacArthur et al. 1993). The alignment, model building and evaluation steps were iterated until a satisfactory result was obtained.

The same procedure was applied to the modelling of the LDLRA domain using the crystal structure 1jja of the human low-density lipoprotein receptor. The LDL-receptor class A domain contains 6 disulphide-bound cysteins (Bieri, Djordjevic et al. 1995). Particular care was therefore taken in correctly aligning the cysteins. A highly conserved cluster of negatively charged amino acids forms a Ca^{2+} coordination site (Daly, Scanlon et al. 1995). The LDLRA repeat has been shown to consist of a beta-hairpin structure followed by a series of beta turns. Based on the pattern of cystein residues, and on the homology detected by template identification tools, the TMPRSS3 sequence appears to contain only one occurrence of the repeat. The quality assessment for the final models are satisfactory and essential structural features such as the disulfide bridge pattern and the Ca^{2+} coordination site of the LDLRA domain appear to be correctly retained. Regions with slightly positive energy value reported for ANOLEA and GROMOS correspond to residues where the alignment is less conserved and gaps in the alignment require *ab-initio* loop remodelling.

The presented examples demonstrate the usefulness and the capabilities of the integrated web-based modelling infrastructure SWISS-MODEL Workspace. In combination with DeepView (Swiss-PdbViewer), it complements the server pipeline and repository (Kopp and Schwede 2004) as powerful and easy to use web-based resources for comparative protein modelling.

ACKNOWLEDGEMENTS

We are grateful to Karol Miaskiewicz, Jack Collins and Robert W. Lebherz (Advanced Biomedical Computing Center, NCIFCRF Frederick, MD, USA) for excellent computational and technical support. We are deeply indebted to Manuel C. Peitsch (Novartis AG, Basel) and to Nicolas Guex (GSK, Raleigh, NC, USA.) for their pioneering work on large-scale protein structure modelling. We would like to acknowledge financial support by the Swiss National Science Foundation (SNF).

REFERENCES

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

Andreeva, A., Howorth, D., et al. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**, D226-229.

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

Bateman, A., Coin, L., et al. (2004). The Pfam protein families database. *Nucleic Acids Res* **32**, D138-141.

Bates, P. A., Kelley, L. A., et al. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* **Suppl 5**, 39-46.

Bieri, S., Djordjevic, J. T., et al. (1995). Disulfide bridges of a cysteine-rich repeat of the LDL receptor ligand-binding domain. *Biochemistry* **34**, 13059-13065.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* **357**, 543-544.

Daly, N. L., Scanlon, M. J., et al. (1995). Three-dimensional structure of a cysteine-rich repeat from the low-density lipoprotein receptor. *Proc Natl Acad Sci U S A* **92**, 6334-6338.

Davies, T. G., Bentley, J., et al. (2002). Structure-based design of a potent purine-based cyclin-dependent kinase inhibitor. *Nat Struct Biol* **9**, 745-749.

Dunker, A. K. and Obradovic, Z. (2001). The protein trinity--linking function and disorder. *Nat Biotechnol* **19**, 805-806.

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

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

Henrick, K. and Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem Sci* **23**, 358-361.

Herter, S., Piper, D. E., et al. (2005). Hepatocyte growth factor is a preferred in vitro substrate for human hepsin, a membrane-anchored serine protease implicated in prostate and ovarian cancers. *Biochem J* **390**, 125-136.

Hillisch A, Pineda LF, Hilgenfeld R (2004). Utility of homology models in the drug discovery process. *Drug Discov Today* **9**, 659-669.

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

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

Hutchinson, E. G. and Thornton, J. M. (1996). PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci* **5**, 212-220.

Iakoucheva, L. M., Brown, C. J., et al. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* **323**, 573-584.

Jeffrey, P. D., Russo, A. A., et al. (1995). Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature* **376**, 313-320.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195-202.

Jones, D. T. and Ward, J. J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **53 Suppl 6**, 573-578.

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

Karplus, K., Barrett, C., et al. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846-856.

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

Kopp, J. and Schwede, T. (2004). Automated protein structure homology modeling: a progress report. *Pharmacogenomics* **5**, 405-416.

Kopp, J. and Schwede, T. (2004). The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res* **32**, D230-234.

Lambert, C., Leonard, N., et al. (2002). ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics* **18**, 1250-1256.

Laskowski, R. A., Chistyakov, V. V., et al. (2005). PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res* **33**, D266-268.

Laskowski, R. A., MacArthur, M. W., et al. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* **26**, 283-291.

Lee, Y. J., Park, D., et al. (2003). Pathogenic mutations but not polymorphisms in congenital and childhood onset autosomal recessive deafness disrupt the proteolytic activity of TMPRSS3. *J Med Genet* **40**, 629-631.

Li, W., Jaroszewski, L., et al. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282-283.

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

Liu, D., Matzuk, M. M., et al. (1998). Cyclin A1 is required for meiosis in the male mouse. *Nat Genet* **20**, 377-380.

Lund, O., Frimand, K., et al. (1997). Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng* **10**, 1241-1248.

Marti-Renom, M. A., Stuart, A. C., et al. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29**, 291-325.

Masmoudi, S., Antonarakis, S. E., et al. (2001). Novel missense mutations of TMPRSS3 in two consanguineous Tunisian families with non-syndromic autosomal recessive deafness. *Hum Mutat* **18**, 101-108.

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

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

Mulder, N. J., Apweiler, R., et al. (2005). InterPro, progress and status in 2005. *Nucleic Acids Res* **33 Database Issue**, D201-205.

Pearl, F., Todd, A., 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**, D247-251.

Peitsch, M. C. (1995). Protein modelling by E-Mail. *BioTechnology* **13**, 658-660.

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

Rychlewski, L. and Fischer, D. (2005). LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci* **14**, 240-5.

Schwede, T., Kopp, J., et al. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* **31**, 3381-3385.

Scott, H. S., Kudoh, J., et al. (2001). Insertion of beta-satellite repeats identifies a transmembrane protease causing both congenital and childhood onset autosomal recessive deafness. *Nat Genet* **27**, 59-63.

Shindyalov, I. N. and Bourne, P. E. (2000). Improving alignments in HM protocol with intermediate sequences. *Forth meeting on the critical assessment of techniques for protein structure prediction* **A**, 92.

Somoza, J. R., Ho, J. D., et al. (2003). The structure of the extracellular region of human hepsin reveals a serine protease domain and a novel scavenger receptor cysteine-rich (SRCR) domain. *Structure (Camb)* **11**, 1123-1131.

Tramontano, A. and Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins* **53 Suppl 6**, 352-368.

Tress, M., Ezkurdia, I., et al. (2005). Assessment of predictions submitted for the CASP6 comparative modelling category. *Proteins*.

van Gunsteren, W. F., Billeter, S. R., et al. (1996). Biomolecular Simulations: The GROMOS96 Manual and User Guide. Zürich, VdF Hochschulverlag ETHZ.

Velankar, S., McNeil, P., et al. (2005). E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res* **33 Database Issue**, D262-265.

Wattenhofer, M., Di Iorio, M. V., et al. (2002). Mutations in the TMPRSS3 gene are a rare cause of childhood nonsyndromic deafness in Caucasian patients. *J Mol Med* **80**, 124-131.

Westbrook, J., Feng, Z., et al. (2003). The Protein Data Bank and structural genomics. *Nucleic Acids Res* **31**, 489-491.

Wheeler, D. L., Barrett, T., et al. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **33 Database Issue**, D39-45.

Wolgemuth, D. J., Lele, K. M., et al. (2004). The A-type cyclins and the meiotic cell cycle in mammalian male germ cells. *Int J Androl* **27**, 192-199.

Zdobnov, E. M. and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848.

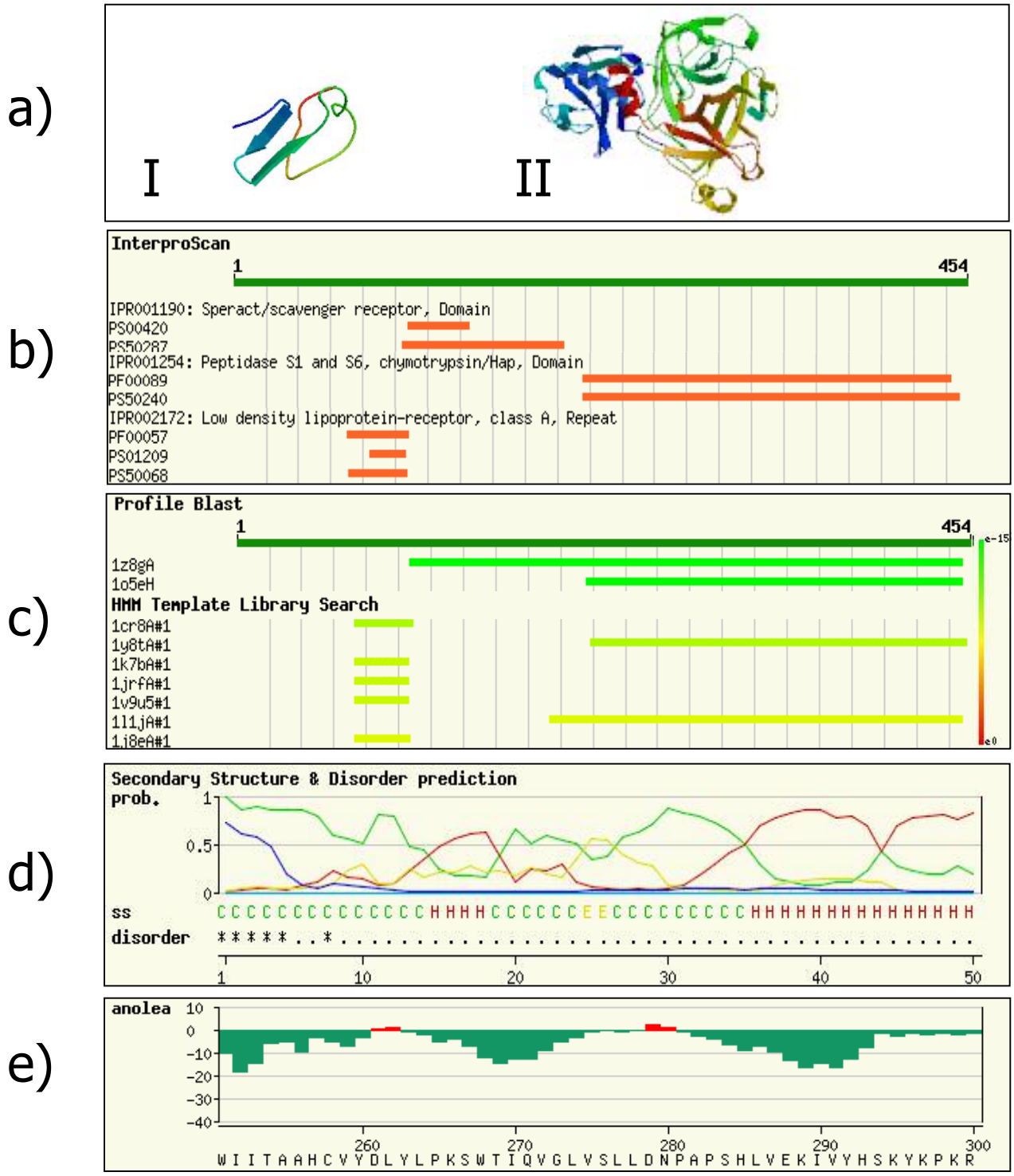

Fig. 1: Graphical output of SWISS-MODEL workspace representing typical steps of a modelling experiment. a) Ribbon representation of three domains modelled for the target protein human transmembrane protease 3 (TMPRSS3): I) LDL receptor domain and II) Scavenger Receptor in complex with the protease domain. b) IprScan of the target sequence detected three domains in the target protein. c) Sequence based searches of the template library identified two segments with suitable template structures. d) Secondary structure and disorder prediction of the target protein. e) Anolea mean force potential plot allows for quality assessment of the final models.