SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling

Article *in* Electrophoresis · December 1997

Impact Factor: 3.03 · DOI: 10.1002/elps.1150181505 · Source: PubMed

2 authors, including:

Manuel C Peitsch
Philip Morris International
263 PUBLICATIONS **19,680** CITATIONS

SEE PROFILE

Nicolas Guex
Manuel C. Peitsch

Geneva Biomedical Research Institute, Glaxo Wellcome Research and Development, Plan-les-Ouates/ Geneva, Switzerland

SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling

Comparative protein modeling is increasingly gaining interest since it is of great assistance during the rational design of mutagenesis experiments. The availability of this method, and the resulting models, has however been restricted by the availability of expensive computer hardware and software. To overcome these limitations, we have developed an environment for comparative protein modeling that consists of SWISS-MODEL, a server for automated comparative protein modeling and of the SWISS-PdbViewer, a sequence to structure workbench. The Swiss-PdbViewer not only acts as a client for SWISS-MODEL, but also provides a large selection of structure analysis and display tools. In addition, we provide the SWISS-MODEL Repository, a database containing more than 3500 automatically generated protein models. By making such tools freely available to the scientific community, we hope to increase the use of protein structures and models in the process of experiment design.

1 Introduction

Understanding protein function often requires well-designed site-directed mutagenesis experiments. These experiments are aimed at the elucidation of the role played by individual residues, during enzyme-based reactions, in molecular recognition or protein structure stabilization. The planning of mutagenesis experiments can be greatly rationalized by insights into the structure of the molecules involved in the studied process. However, collecting experimental data allowing the elucidation of protein structure is not an easy enterprise since experimental structure determination is often hampered by difficulties in obtaining sufficient pure protein, diffracting crystals and many other technical aspects and time constraints. Therefore the gap between known three-dimensional (3-D) structures and available amino acid sequences is growing exponentially. Indeed, the rate of sequence determination is much higher than the rate of structure elucidation and is illustrated by the fact that the SWISS-PROT/TREMBL databases [1] contain more than 150 000 sequences while the Brookhaven Protein Data Bank (PDB) [2] contains only 9000 protein chains, approximately 5000 of which have distinct sequences. This gap will increase even further with the explosion of large-scale genome sequencing projects. In this context, it is not surprising that predictive methods to derive the 3-D structure of a protein are rapidly gaining interest.

Protein modeling requires profound knowledge and understanding of the rules underlying protein structure, expensive hardware and software, as well as expert knowledge in their manipulation. This combination is not generally available in molecular biology laboratories. Therefore only a limited number of scientists uses these approaches while designing site-directed mutagenesis experiments. To overcome these limitations we have developed SWISS-MODEL (Table 1), an automated comparative protein modeling server [3, 4] and the SWISS-MODEL Repository [5], a database of automatically generated protein models which can be readily downloaded.

Correspondence: Dr. Manuel C. Peitsch, Geneva Biomedical Research Institute, 14, chemin des Aulx, 1228 Plan-les-Ouates/Geneva, Switzerland (Tel: +41-22-706-9920; Fax: +41-22-794-6965; E-mail: mcp 13936@ggr.co.uk)

Nonstandard abbreviations: PDB, Brookhaven Protein Data Bank; **rmsd,** relative mean square deviation

Keywords: Protein modeling / Protein structure / Database

2 SWISS-MODEL and Swiss-PdbViewer

The SWISS-MODEL server is mainly used in a fully automated mode where no user input, other than the sequence of the protein to model, is required. This is of course the easiest and most user-friendly way to obtain a protein model. As we will show later, however, it would often be suitable to use the server in the more advanced mode and take full advantage of its capabilities. This is more tedious, since the user must provide an alignment and a command file in a specified syntax. We have addressed this problem by developing the SWISS-PdbViewer, a graphical front-end to SWISS-MODEL, for both Macintosh and PCs. The Swiss-PdbViewer not only provides advanced molecular display features and real-time visual feedback during structure modeling, but also features direct submission to the SWISS-MODEL model server.

2.1 The SWISS-MODEL server

Proteins from different origins can have very similar sequences, and it is generally accepted that high sequence similarity is reflected by distinct structure similarity. Indeed, the relative mean square deviation (rmsd) of the Cα coordinates for protein cores sharing 50% residue identity is expected to be around 1 Å [6]. This fact served as the premise for the development of comparative protein modeling methods, which consists of the extrapolation of the structure for a new (target) sequence from the known 3-D structure of related family members (templates) (for review see [7]).

The comparative protein modeling process was automated and implemented in the Internet server SWISS-

0173-0835/97/1515-2714 $17.50+.50/0

MODEL [3, 4]. The process begins with the identification of suitable template structures based on their sequence similarity with the target sequence. This is achieved by searching a database of sequences with known 3-D structure using the sequence alignment tools BLAST [8] and FASTA [9]. A structurally corrected multiple sequence alignment of the templates is then generated using the 3-D superposition module of ProMod (Protein Modeling tool) [4]. The target sequence is then added to the above multiple alignment using SIM [10]. This final alignment defines the spatial correspondence of each target and template residue and serves as the guide for the model building procedure. The generation of model coordinates is automated in ProMod [4] and follows these steps: (i) the construction of an averaged framework [11] from the superimposed template structures; (ii) the generation of atomic coordinates, derived from the averaged framework, using the multiple sequence alignment described above; (iii) the rebuilding of nonconserved loops (including both insertions and deletions) from their "stems" by structural homology searches through the Brookhaven Data Bank as previously described [12, 13]; (iv) the completion of the main chain using a library of backbone elements (pentapeptides) derived from the best X-ray structures (< 2 Å resolution), and (v) the reconstitution of lacking side chains and the correction of existing ones using a library of allowed rotamers [14]. The quality of the model is assessed by computing its 3D-1D profile [15] and using Prosa II [16] (see the note on model accuracy). Optimization of bond geometry and relief of unfavorable nonbonded contacts is then performed by 50 steps of steepest descent followed by 500 steps of conjugate gradient energy minimization using the force field package CHARMM [17] with the PARAM22 parameter set. A model confidence factor [4], describing the degree of uncertainty linked to each residue, is computed during the modeling procedure and occupies the crystallographic B-factor field in the final coordinate file. The SWISS-MODEL server presently offers three modes of function: the First Approach Mode, the Optimize Mode and the GPCR-Modeling Mode.

2.1.1 First Approach Mode

The First Approach Mode only requires the submission of a raw amino acid sequence or its SWISS-PROT ID (or AC) code, but can also accept a list of pre-selected template structures. In this mode, SWISS-MODEL will go through the complete procedure described above. The final model, its 3D-1D profile [15] and the Prosa II [16] analysis results will be returned to the user via e-mail. In order to use the Optimize Mode (see below) at a later stage, the user must request the verbose output, which will cause both the used sequence alignment and the ProMod command file to be returned to the user. These can be altered and resubmitted through the Optimize Mode.

2.1.2 Optimize Mode

The First Approach Mode generally returns good-quality models, which differ by approximately 1 Å rmsd from control experimental structures (as calculated using Cα atoms) for sequence identity levels around 50%. As expected, this degree of structural similarity corresponds to that described by Chothia and Lesk [6] when comparing experimentally determined structures sharing the same sequence identity levels. However, automated sequence alignment programs are often unable to accurately position insertions and deletions when the local levels of sequence similarity fall below 35%. In such cases, one can optimize the model by manually altering the alignment in these regions. Until now this process was tedious, not only because of the difficulty to refine an alignment without interactive feedback, but also due to the syntax constraints imposed by the server's Optimize Mode. The Swiss-PdbViewer now provides a front-end to SWISS-MODEL and allows the user to directly submit custom modeling requests *via* the Optimize Mode. Hence, sequences sharing very low levels of identity, which are rejected by the First Approach Mode, can now be modeled easily.

2.1.3 GPCR-Modeling Mode

Finally, the GPCR-Modeling Mode uses predefined GCPR modeling templates [18] which were derived using a rule-based approach developed by P. Herzyk and R. Hubbard [19].

2.2 The Swiss-PdbViewer

2.2.1 Interface

Swiss-PdbViewer [20] (Table 1) was designed to maximize interactivity and is aimed at experimental scientists who do not have the time to deal with complex computer commands. One of the Swiss-PdbViewer's strength is its powerful interface, which allows to quickly "navigate" within protein structures, easily alter the display of

Table 1. A few relevant Internet locations

Service	Adress
ExPASy molecular biology server	http://www.expasy.ch/
SWISS-MODEL	http://www.expasy.ch/swissmod/SWISS-MODEL.html
SWISS-MODEL Repository	http://www.expasy.ch/swissmod/swmr-top.html
Swiss-PdbViewer	http://www.expasy.ch/spdbv/mainpage.html http://www.pdb.bnl.gov/expasy/spdbv/mainpage.html
SWISS-3DIMAGE	http://www.expasy.ch/sw3d/sw3d-top.html
SWISS-PROT	http://www.expasy.ch/sprot/sprot-top.html
SWISS-2DPAGE	http://www.expasy.ch/ch2d/ch2d-top.html
Brookhaven Protein Data Bank (PDB)	http://www.pdb.bnl.gov/
POV-Ray	http://www.povray.org/
Quickdraw™ 3D	http://www.quickdraw3d.com/

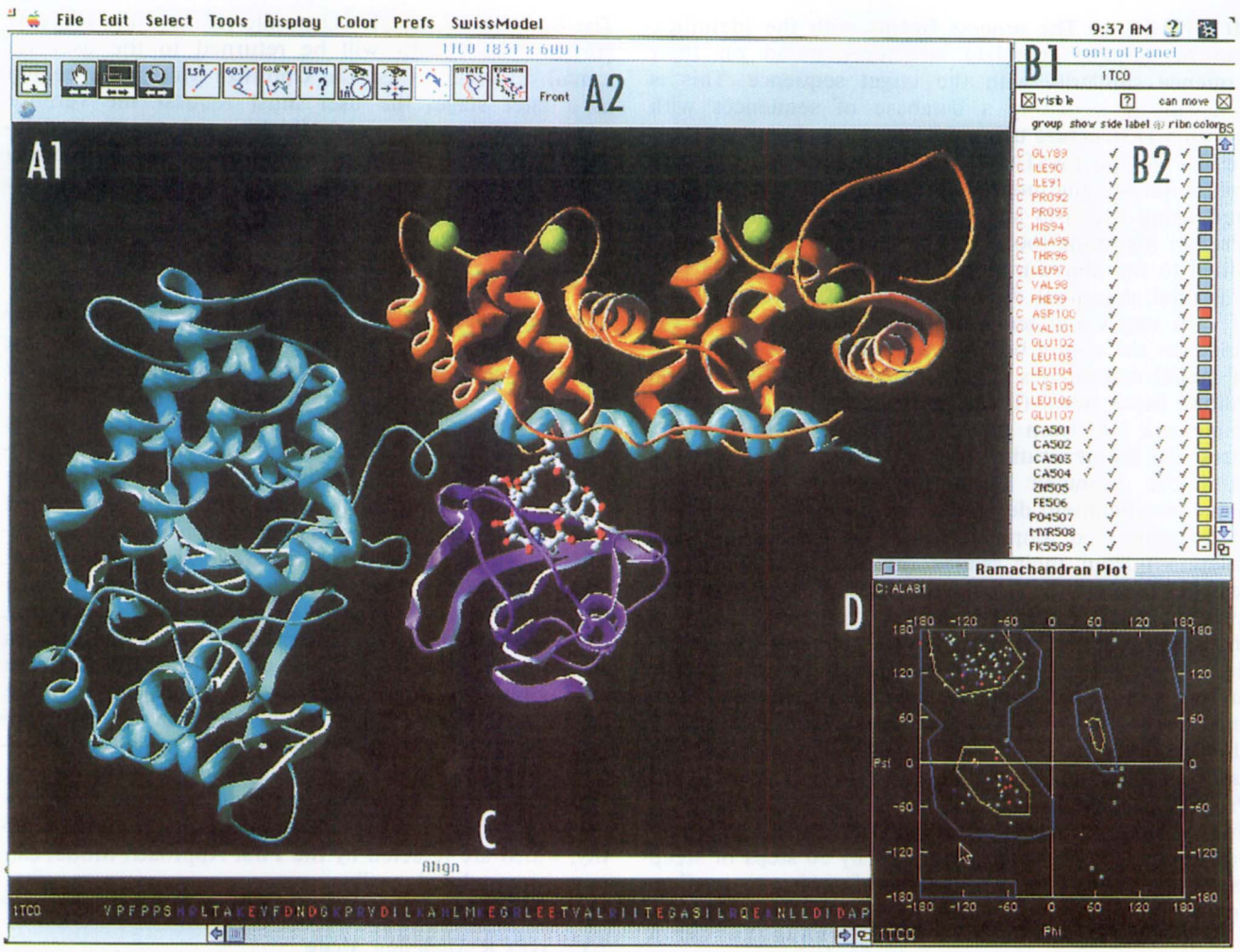

Figure 1. The Swiss-PdbViewer workspace. (A1) This is the main window, in which molecules are displayed (current structure 1TCO). Although the default mode is "wireframe", solid images can be rendered through Quickdraw™ 3D. (A2) Buttons for the commonly used tools; from left to right: move, zoom, rotate, measure distances, angles, dihedral angles, view neighbors groups, center view, fit layers, mutate side chains, adjust dihedral angles. (B1) This is the Control Panel's header, allowing us to switch among loaded proteins. (B2) List of groups (residues) for the currently active molecule, with individual controls allowing easy alteration of the display status of each residue. (C) Alignment window, for an easy view of structural alignments, residue properties, rms deviation to reference stucture. This window also allows manual adjustments of the sequence alignment during comparative modeling. (D) Ramachandran plots of selected residues of the currently active layer. More tools are availabe from the menus.

selected parts of the protein and measure distances, angles, and dihedral angles between atoms. Several files can be loaded simultaneously. Each loaded file appears in a separate layer, which in turn holds individual chains, themselves composed of groups. These groups are either amino acids, nucleotides or heterogeneous groups (*e.g.* NAD, HEME). The first loaded coordinate set will be considered the reference structure. The Swiss-PdbViewer assigns Kollman's atom types [21], which are used for H-bond detection.

The workspace is divided into three major windows (Fig. 1). The main window is used to display the molecules and provides a palette of buttons to access the most commonly used tools. A slab mode, which allows us to see only a "slice of the protein", can be set to a user-defined thickness and can be moved along the *z*-axis. As soon as the Cα atom of the residue is visible, the whole side chain is shown, preventing disconnected atoms from being displayed. A second window acts as a control panel, of which the upper part allows the selection of the current active layer. The lower part of the window provides fine display controls for the individual groups of that layer. A third window allows visualizing of the structural alignments, and is also used to manually alter alignment before issuing modeling requests to SWISS-MODEL. This window also provides a visual feedback of the spatial location of a sequence position: pointing at a residue in the alignment causes the amino acid to flash in the main window. Furthermore, the residue identification and its rmsd to the reference structure is shown at the lower left corner of the alignment window. A fourth window providing a Ramachandran plot of the selected groups of the current active layer can be opened on request. Residue selections are automatically reflected in all windows.

2.2.2 Rendering

The molecules are displayed as "wireframes" per default. High quality images can be generated from the current

view, including all visible groups, through Apple™ QuickDraw™ 3D (Table 1), which is freely available on PowerMacs™, Windows NT™ and Windows 95™. The advantage of using such libraries is that this will allow transparent use of hardware acceleration as well as multi-processor capabilities. Furthermore, scenes for the popular free ray-tracer POV-Ray (Table 1) can be generated, and all color preferences as well as light positions are taken directly from the QuickDraw™ 3D user preferences. No editing, apart from moving the camera along the *z*-axis, is required to get exactly the same view with POV-Ray as with QuickDraw™ 3D.

2.2.3 H-bond detection

Case 1: explicit hydrogen positions. The detection of H-bonds is based on the relative donor, hydrogen, and acceptor atoms with respect to user-provided distances and angles.
Case 2: no hydrogen atoms. The detection of putative H-bonds is based on user-provided distances separating donor and acceptor atoms. Additionally, a loose cut-off angle can be provided to further limit the detection to only "reasonable" possibilities.

2.2.4 Ramachandran plots

The Ramachandran plot is subdivided into core and allowed zones, *etc.*, as defined by Morris and colleagues [22]. Amino acids out of the core and the allowed zones, or bearing unusual Ω angles, can be selected for easy structure inspection. Moving the cursor over any group of points indicates the list of residues in that group. A tabulated text file containing residue name, secondary structure assignment, as well as Ω, Φ and Ψ angles of the selected amino acids can be exported.

2.2.5 Secondary structure detection

The secondary structure is evaluated in two passes. The first pass detects ranges of amino acids bearing Φ/Ψ angles compatible with an α-helix conformation. A second pass extends these nucleation sites using H-bonding patterns. Only helices of three or more residues are kept. Remaining residues are checked for parallel and anti-parallel β-strands using H-bond patterns.

2.2.6 Superposing proteins

Comparing molecular structures to highlight their differences is one of the most common tasks in structural biology. Superposition was traditionally achieved by picking three or more corresponding atoms in each molecule. Although the Swiss-PdbViewer offers this feature, it is tedious and does not guarantee an optimal result. Therefore, two other methods are provided. The first consists in selecting residues that are known to be equivalent in the two proteins (*e.g.* active site residues) and then selecting the "Fit Molecules" function. The two proteins will be centered at the centroid of the selected residues, and rotated until the best match is achieved. The user can select Cα or backbone atoms (or even complete side chains if they are identical in both proteins) to be used during the calculations. The rmsd for the involved atoms will be automatically computed and displayed. The second superposition method will automatically detect conserved amino acids to use during the process. This is achieved by aligning the sequences using SIM [10], and selecting pairs of strongly conserved residues from the best local alignment. This last method only works, however, if the proteins to be superposed share sufficient similarity.

Once the proteins are superposed by either of these methods, it is possible to deduce a structural alignment. This is achieved by the identification of backbone portions with low rmsd, which are used as nucleation starts. These are then extended up to a point where the structures are too divergent. Finally, gaps are added to accommodate insertions and deletions. These alignments can be saved as text files, with various options, including one allowing the addition of the secondary structure assignment below each amino acid. Residues can then be colored according to their rmsd from the reference structure, which allows a quick visual inspection of the structural conservation. Corresponding amino acids with a backbone rmsd below 0.2 Å are colored in dark blue whereas residues with an rmsd above 5 Å are colored in red. A linear color gradient from blue to green to yellow to red is applied in-between.

2.2.7 Mutations

Swiss-PdbViewer provides an amino acids mutation utility. The selected side chain can be replaced by any amino acid rotamer available in the provided library which is identical to the one used by SWISS-MODEL [4]. During this process, the rotamer that best fits into the local environment is automatically selected in order to minimize bad contacts and maximize possible H-bonds. The user can, however, easily browse the library to select another possible rotamer. The side chain torsion angles can be further adjusted manually. The display of bad contacts and H-bonds is updated in real time while the side chain atom names are updated to match the IUPAC [23] nomenclature.

2.2.8 Comparative protein modeling

As the number of experimentally determined 3-D structures has dramatically increased over the past years, comparative modeling has become a convenient way to obtain structural information about proteins for which no experimental structure is available. This method however, is restricted to sequences sharing significant similarity with known structures. As outlined above, the quality of a model greatly depends on the alignment provided. The Swiss-PdbViewer allows the generatation of custom alignments which can be submitted to SWISS-MODEL. The modeling templates are provided by the PDB, which raises two problems: First, many PDB entries contain more than one chain, and second, some amino acids are incomplete, missing from the entry, or the primary sequence simply differs from what can be

found in SWISS-PROT. In order to circumvent these problems we prepared the ExPDB structure database, which is composed of separated chains with renumbered residues (starting at 1) and excludes some of the old entries which lack sequence information. Each ExPDB entry code is composed of the chain number – consecutive numbering as they appear in the PDB file – followed by the genuine PDB code. For example, the PDB entry 1BMD holds the two protein chains A and B, which translate into the two ExPDB entries 11BMD and *21BMD*. Furthermore, we generated the ExPDB sequence database which only holds the renumbered chains. The ExPDB sequence database is used for template identification.

A typical comparative modeling project will follow these steps: (i) Load the protein to model as a raw sequence (target sequence). (ii) Look for suitable modeling templates, which is achieved by comparing the target sequence with all entries of the ExPDB sequence database using blastp [8]. This can be done from the Swiss-PdbViewer and requires a Web browser as a helper application. If suitable templates are available, they will be sorted by statistical significance and can be readily downloaded. Alternately, if the user already knows the template's entry code, he can download it directly. (iii) If several template structures are selected, superpose them and generate the structural alignment. (iv) Align the sequence to model with the templates. An initial alignment can be obtained with the "Magic Fit" tool, and then improved by manual alterations. This process is eased by real-time feedback and property-based amino acid coloring. One should take care to avoid placing gaps within secondary structure elements and minimize the number of exposed nonpolar residues (for soluble proteins). (v) Decide which residues of the templates should be ignored during model building by marking them with a "*" within the Swiss-PdbViewer. (vi) Submit the alignment to SWISS-MODEL. A Web browser will be launched by the Swiss-PdbViewer with the Optimize Mode submission form containing all the necessary information. Just hit the "Submit" button and wait for the model coordinates to be returned via e-mail.

3 Large-scale protein modeling and the SWISS-MODEL Repository

3.1 Large scale proteins modeling

There is no doubt that the scrutiny of multiple sequence alignments is more informative than the analysis of an isolated sequence. This allows the identification of conserved and variable residues, and improves the rationalization of site-directed mutagenesis experiments. Likewise, comparative structure analysis involving several members of a protein family, as opposed to single structure inspection, is expected to be invaluable for the understanding of their functional differences and for rational combinatorial library design. It will thus be increasingly valuable to have models for as many members of a protein family as possible. By comparing all entries of the SWISS-PROT database (release 34) with a nonredundant subset of all sequences of known 3-D structure we found that approximately 15% of the protein sequences have at least one suitable modeling template. This number increases every year by roughly 5%, showing that the growth in available templates allows an ever-increasing number of sequences to be modeled by comparative methods. Hence, we might speculate that comparative modeling procedures could be applied to almost 30% of the proteins by the beginning of the next millennium. For these sequences it is generally feasible to attempt comparative protein modeling using a completely automated approach. However, large multi-domain complex models can presently not be built, mainly because no suitable modeling templates are available and because the prediction of protein-protein contacts is inaccurate and still in its infancy.

3.2 The SWISS-MODEL Repository

We have taken a species-based approach to large-scale protein modeling and submitted – in batch mode – all known members of the proteomes of *Escherichia coli* [5], *Haemophilus influenzae, Mycoplasma genitalium, Mycobacterium tuberculosis*, and *Basillus subtilis* to the SWISS-MODEL server. In this fully automated approach we required at least 35% residue identity within the segments aligned to the modeling template. This ensures that the quality of the resulting models does not fall below an acceptable threshold. All resulting protein models are annotated with information regarding the modeling template(s) as well as the alignment between template and target sequences. The connectivity with other databases is provided by the same identification and accession codes as the corresponding SWISS-PROT entry. The models are stored as individual files and can be accessed through the SWISS-MODEL Web pages (Table 1). In addition, every time a SWISS-PROT or a SWISS-2DPAGE entry is requested from the ExPASy (Expert Protein Analysis System) Web server (Table 1) [24], a hyperlink to the coordinate file is provided if a corresponding model exists in the SWISS-MODEL Repository. Model coordinates can be readily downloaded and imported into the Swiss-PdbViewer ([20], this paper).

4 The future

(i) Increased flexibility: One can already use the Swiss-PdbViewer to position different modeling templates relative to each other. These positions will be transmitted to SWISS-MODEL, along with the alignment. Thereby we will provide multi-domain protein building capabilities through SWISS-MODEL. Partial models generated with the Swiss-PdbViewer for specific reasons (*e.g.*, custom backbone and loops) will be accepted and completed by SWISS-MODEL. (ii) Loop building methods: New approaches to build protein loops are currently being explored. They will be included into future versions of the Swiss-PdbViewer and SWISS-MODEL. (iii) Force field and energy minimization: The Swiss-PdbViewer and SWISS-MODEL do not provide an energy minimization tool which can be used independently of a complete modeling process. This means that the local refinement of a mutation or loop generated within the Swiss-

PdbViewer cannot be achieved. To overcome this limitation, we plan to add limited local minimization capabilities to the Swiss-PdbViewer. (iv) Empirical energy potentials (potentials of mean force): Potentials of mean force have gained much attention during the last years, and have proven most useful in assessing the sequence to structure fitness of a model. This translates into a model quality judgement [7]. We are currently developing a complete potential derivation environment as a new feature of the Swiss-PdbViewer. These potentials can be used during a modeling project to evaluate the "value" of a sequence alignment. The total energy as well as local energies are computed and their display is updated in real time during sequence alignment alterations. (v) Work environment: A future version of the Swiss-PdbViewer will include a work environment conservation feature. This will allow users to resume a project exactly where they left it.

5 Example

5.1 Possibilities and limitations

In order to illustrate the possibilities and limitations of comparative protein modeling, we have built a model for porcine cytoplasmic malate dehydrogenase (PDB entry 4MDH) [25] using the *Thermus flavus* malate dehydrogenase (PDB entry 1BMD chain A – ExPDB entry 11BMD) [26] as a template. The project was carried out solely with the above described tools. The sequence of 4MDH (SWISS-PROT accession code P11708) was loaded, from the SWISS-MODEL menu, into the Swiss-PdbViewer, and appeared as a single long helix. From the same menu, we performed a search for suitable modeling templates. We ignored the best hit, which obviously was the control structure 4MDH, and selected the second hit (11BMD), sharing 54% identity with the target sequence for the modeling procedure. The coordinates of 11BMD were loaded into the Swiss-PdbViewer. The target sequence was then aligned automatically with the template sequence using SIM [10] (Fig. 2) and an initial model was built (model 1) by submitting the alignment to SWISS-MODEL from the Swiss-PdbViewer. The automatically created alignment was inspected carefully. We observed that the gap introduced after 4MDH:W257 is immediately compensated by another one just before 11BMD:W251 (Fig. 2). These two gaps were obviously introduced to satisfy the alignment of tryptophans in the two structures. Indeed, tryptophan is a rare amino acid and has the highest conservation weight in the PAM200 matrix. After removal of this alignment artifact, a second model (model 2) was built.

The remaining parts of the alignment were also inspected, and particular attention was devoted to the 4MDH region spanning H196 to F222 (H196 to F217 in 11BMD, respectively), where 4MDH contains five more amino acids than 11BMD, leading to two gaps (Fig. 2): The first gap is the insertion of the 4MDH residues KEVG (204 to 207) between P203 and A204 of 11BMD. Residues inserted in this location would point towards the core of the protein (Fig. 3A). Therefore we altered the alignment interactively in order to slide the gap to the tip of the nearby loop. As a result, the 4MDH sequence QAKE (202 to 205) will be inserted between G201 and R202 of 11BMD (Fig. 3B). The second gap places 4MDH:S216 between E211 and W212 of 11BMD (Fig. 3A). This is within an α-helix, and would disrupt this secondary structure element. Therefore, it is better to move the insertion point just before the α-helix, by placing 4MDH:D214 between D209 and M210 of 11BMD (Fig. 3B). A third model (model 3) was then built, based on this new alignment. The common core of the template (11BMD) and experimental control (4MDH) structures contains 297 residues and the Cα atoms of both structures diverge by 1.1 Å. The same value was obtained when comparing model 3 with the control structure (Table 2). These values are expected for a sequence identity level of 54% [6].

The analysis of the resulting coordinate sets shows that model 1 and model 2 are almost identical (rmsd 0.08 Å for 2544 atoms). Nevertheless, SWISS-MODEL followed two different paths during the construction of the region spanning the corrected one-residue gap (4MDH:W257). In model 1, the region was rebuilt using the spare-part algorithm, while for model 2, the coordinates were derived from the template during framework construction. Both models are equally correct in this region since they diverge only slightly from the 4MDH control structure. From a sequence alignment point of view it seemed logical that SIM aligned the two tryptophans (see above); however, the constraints imposed on the local structure do not allow their aligning. Indeed, the relative spatial positions of the tryptophans differs by one residue in 4MDH and 11BMD.

The two other adjustments made to the alignment had a far greater impact on the model structure as evidenced by the comparison of model 1 with model 3 (Table 3). The structure of the affected region was greatly improved (Fig. 4) as demonstrated by a drop in all atom rmsd from 6.19 Å to 2.84 Å (Table 3). This illustrates the

```
4MDH.AA    1  SEPIRVLVTG AAGQIAYSLL YSIGNGSVFG KDQPIILVLL DITPMMGVLD
11BMD      1  KAPVRVAVTG AAGQIGYSLL FRIAAGEMLG KDQPVILQLL EIPQAMKALE
                *.** *** *****.**** . *. * ..* ****.** ** .*   * *.

4MDH.AA   51  GVLMELQDCA LPLLKDVIAT DKEEIAFKDL DVAILVGSMP RRDGMERKDL
11BMD     51  GVVMELEDCA FPLLAGLEAT DDPDVAFKDA DYALLVGAAP RKAGMERRDL
              **.***.*** .***  . ** *  ..**** * *.***. * *. ****.**

4MDH.AA  101  LKANVKIFKC QGAALDKYAK KSVKVIVVGN PANTNCLTAS KSAPSIPKEN
11BMD    101  LQVNGKIFTE QGRALAEVAK KDVKVLVVGN PANTNALIAY KNAPGLNPRN
              * * ***   ** **   ** * ***.**** ***** * *  *.**..   *

4MDH.AA  151  FSCLTRLDHN RAKAQIALKL GVTSDDVKNV IIWGNHSSTQ YPDVNHAKVK
11BMD    151  FTAMTRLDHN RAKAQLAKKT GTGVDRIRRM TVWGNHSSTM FPDLFHAEVD
              *. .****** *****.* *  *  * ... .******* .**. ** *

4MDH.AA  201  LQAKEVGVYE AVKDDSWLKG EFITTVQQRG AAVIKARKLS SAMSAAKAIC
11BMD    201  GRP----ALE LVDME-WYEK VFIPTVAQRG AAIIQARGAS SAASAANAAI
               ..      * *  . *    ** ** *** **.* **  * ** ***.*

4MDH.AA  251  DHVRDIW-FG TPEGEFVSMG IISDGNSYGV PDDLLYSFPV TIKDKTWKIV
11BMD    246  EHIRD-WALG TPEGDWVSMA VPSQGE-YGI PEGIVYSFPV TAKDGAYRVV
              .*.** * .* ****. ***. . .*.*. **. *. ..***** * ** . ..*

4MDH.AA  300  EGLPINDFSR EKMDLTAKEL AEEKETAFEF LSSA
11BMD    294  EGLEINEFAR KRMEITAQEL LDEMEQVKAL GLI
              *** **.*.*  .*..** **  .*.*    .
```

Figure 2. Automatic alignment of 4MDH sequence onto the 11BMD template from the ExPDB database (PDB entry 1BMD). This was used to build model 1. Stars and dots are drawn below identical and similar amino acids, respectively.

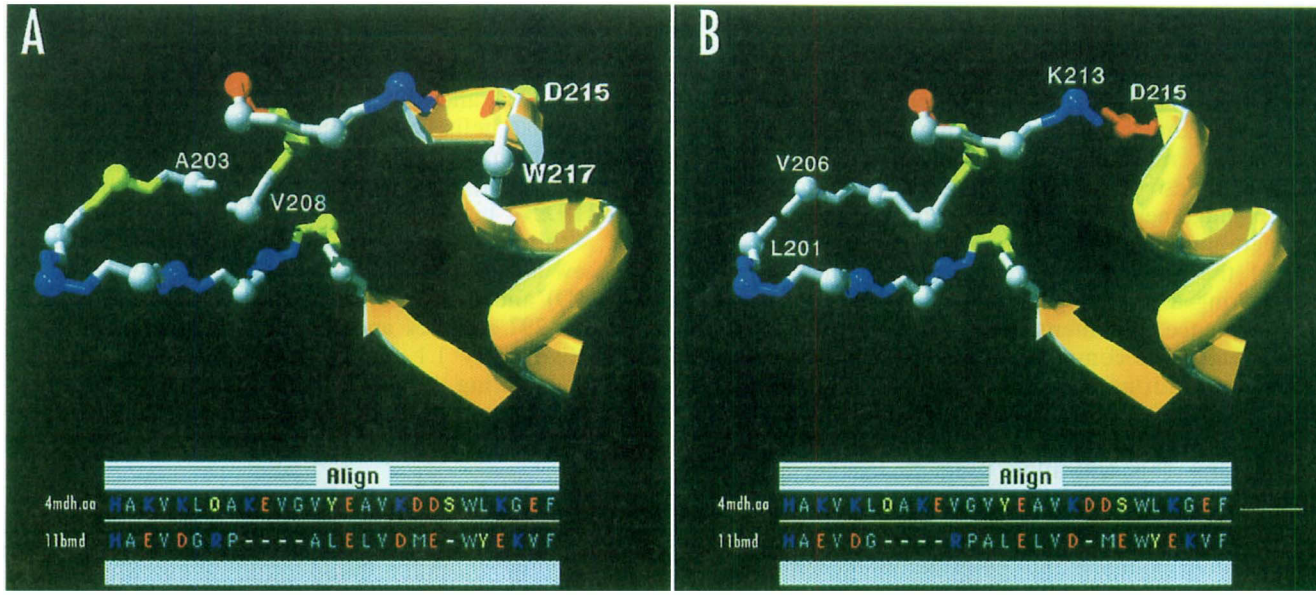

Figure 3. (A) Initial placement of gaps after automatic alignment would require inserting four amino acid towards the protein core (between A203 and V208) and one amino acid within an α-helix (between D215 and W217). (B) Alignment refined manually in order to place the gaps at the extremity of a loop and outside of secondary structure elements. Rendering was performed with POV-Ray 3.01.

Table 2. Relative mean square deviations between the modeling template or the best model and the control structure of malate dehydrogenase

Atoms	Number of atoms[a)]	11BMD[a)]	Model 3[b)]	Number of atoms[c)]	Model 3[c)]
CA	297	1.1	1.1	303	1.1
Backbone	1188	1.1	1.1	1212	1.1
All	2275	—[d)]	1.8	2321	1.8

The experimental control structure is the chain A of the PDB entry 4MDH, while the template is the ExPDB entry 11BMD.

a) Only 11BMD amino acids in the common core – corresponding Cα atom pairs diverging by less than 3 Å – were considered during the calculation. The two main excluded regions are those undergoing a conformational change (see text below and Fig. 5).

b) Only residues corresponding to those used for 11BMD are considered during the calculation.

c) Same as a) but for model 3. The greater number of atoms involved in the calculation shows that some of the modeled loops were accurately predicted. Here again, the two main regions being excluded undergo a conformation change (Fig. 5).

d) As the sequences differ, this value cannot be computed.

importance of manual refinement and their potential effect on model structures. Such refinement, however, will not guarantee a perfect model. Indeed, as mentioned previously, SWISS-MODEL will take the backbone position of the template(s) to build the framework, reconstructing only regions where insertions/deletions occur. In the present case, two regions of the model (M89 to A103 and Q227 to K238) have a higher than average rmsd to the 4MDH control structure, although the template and target sequences are quite similar (Fig. 5). These two regions are in an open conformation in the template used to build the model (11BMD), whereas they are in a closed conformation in the actual 4MDH structure. This displacement results from the fact that 11BMD has been crystallized in presence of reduced NAD (NADH) whereas 4MDH has been crystallized in presence of oxidized NAD (NAD+) and sulfate (the latter approximately occupying the malate binding site). The movement of this loop, which is conserved among MDH and LDH, is well-documented and plays a crucial role during the catalysis (for example the highly conserved 4MDH:R91 can, in the closed conformation, help maintain the malate in a suitable position during the catalysis) [27–31]. As the template used was in an open conformation, the resulting model is also in an open conformation. The model obtained is quite accurate for this form, but the user has to realize that a closed conformation could by no means have been predicted.

5.2 Note on model accuracy

The quality of the model coming back by e-mail greatly depends on the alignment provided to SWISS-MODEL. Careful inspection and manual refinement of the auto-

Table 3. Relative mean square deviations between two models and the experimental control structure of malate dehydrogenase

Atoms	Number of atoms[a)]	Model 1[a)]	Model 3[a)]	Number of atoms[b)]	Model 1[b)]	Model 3[b)]
CA	332	2.7	2.3	27	5.5	2.3
Backbone	1328	3.3	2.3	108	5.5	2.3
All	2544	2.7	2.9	217	6.2	2.8

The experimental control structure is the chain A of the PDB entry 4MDH.

a) All residues, located in the core and the loops, are considered during the calculation.

b) Only the amino acids from H196 to F222 are considered.

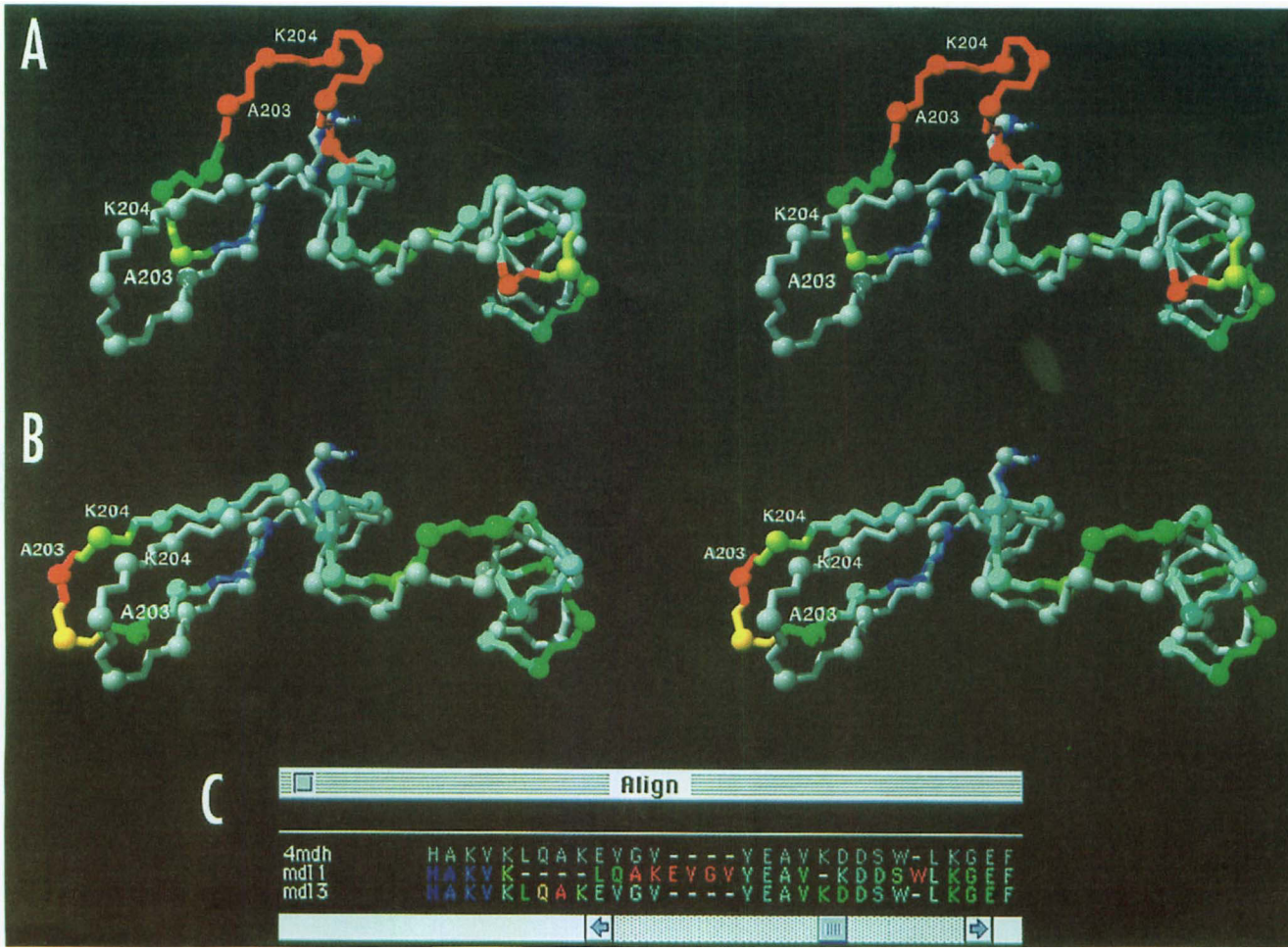

Figure 4. Stereo view of the region ranging from H196 to F222. The 4MDH control structure appears in grey, while model 1 (A) and model 2 (B) are colored by their backbone rmsd to the 4MDH control structure. (C) Alignment summary. The rmsd is color coded from blue to red according to this linear scale: 0.4 Å for dark blue, 1.2 Å for light blue, 2.6 Å for green, 4.3 Å for orange, and > 5 Å for red. Cα atoms are shown as spheres. Rendering was performed with POV-Ray 3.01.

matic alignment before submitting a modeling request is thus quite important and can improve the model quality. However, large conformational changes cannot be predicted as the model backbone is constructed by averaging the template(s) backbone(s). It is important to realize that the technical aspects of comparative protein modeling described above imply that the accuracy of a model is essentially limited by the deviation of the templates relative to the experimental control structure. As a consequence, the Cα atoms of protein models which share 35–50% sequence identity with their templates will generally deviate by 1.0 to 1.5 Å from their experimental counter parts, as do similarly related experimental structures [6]. Furthermore, structural differences between predicted and experimental structures have two sources (i) the errors inherent to the modeling procedures and (ii) the variations caused by the molecular environment and data collection method incorporated into experimentally elucidated structures which will be used as modeling templates. Indeed, crystallographic structures of identical proteins can vary not only because of experimental errors and differences in data collection conditions (illustrated in [32]) and refinement, but also because of different crystal lattice contacts and the presence or absence of ligands (the example herein). One of the most interesting examples in which several structures of the same protein, determined by different methods, were compared involves interleukin-4 (IL-4) [33 and references therein]. This cytokine consists of a 130-residue four-helix bundle, and its structure was elucidated by X-ray crystallography as well as by NMR. The backbones of three IL-4 crystal structures (PDB entries 1RCB, 2INT and 1HIK) show an rmsd of 0.4 to 0.9 Å, while those of three IL-4 NMR forms (PDB entries 1ITM, 1CYL and 2CYK) give rmsd of 1.2 to 2.6 Å. These values illustrate the structural differences due to experimental procedures and the molecular environment at the time of data collection. Therefore, "a protein model derived by comparative methods cannot be more accurate than the difference between the NMR and crystallographic structure of the same protein" [33].

Nonconserved loops are expected to be the least reliable portions of a protein model and often deviate markedly from experimentally determined control structures. In many cases, however, these loops also correspond to the most flexible parts of the protein structure as evidenced by their high cyrstallographic temperature factors (or multiple divergent solutions in NMR experiments). The core side-chains, which are the most conserved in any

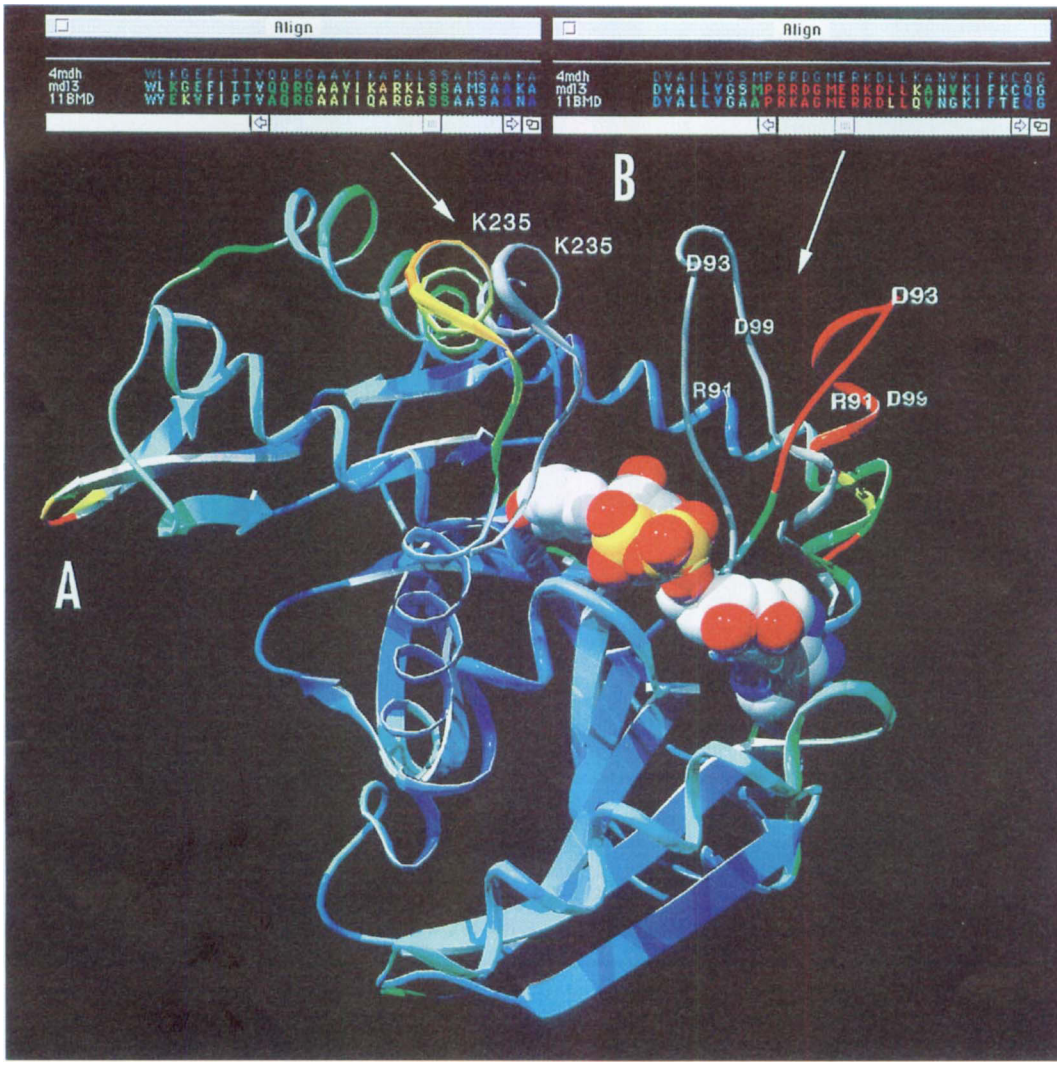

Figure 5. Ribbon representation of model 3 color-coded by its rmsd to the 4MDH control structure. (A) Region enhanced by manual refinement of the alignment. (B) The two regions of the model presenting a high rmsd to the control structure are indicated by arrows and the corresponding alignments are presented. The corresponding regions of the control structure are depicted by grey ribbons. The rest of the control structure is hidden for clarity reasons. The ExPDB entry 11BMD (PDB entry 1BMD) , which was used as the modeling template, is not shown, but its backbone matches quite precisely the one of model 3. The NAD group, taken from the control structure, appears as spheres. Rendering was performed with POV-Ray 3.01.

given protein family, are usually found in approximately the same orientation as in the experimental control structures. In contrast, the more variable surface amino acids will tend to show more deviations since they are generally free to rotate in the solvent.

Some structural aspects of a protein model can be verified using methods based on the inverse folding approach. Two of them, namely the 3D-1D profile verification method [15] and "Prosall" developed by M. Sippl [16], are widely used. The 3D-1D profile of a protein structure is calculated by adding the probability of occurrence for each residue in its 3-D context [15]. Each of the twenty amino acids has a certain probability of being located in any of the environmental classes (defined by criteria such as solvent-accessible surface, buried polar, exposed nonpolar area and secondary structure) defined by Eisenberg and colleagues. In contrast, Prosall [16] relies on empirical energy potentials derived from the pairwise interactions observed in well-defined protein structures. These terms are summed over all residues in a model and result in a more or less favorable energy. Both methods can detect a global sequence to structure incompatibility and more localized errors such as β-strands that are "out of register" or clusters of buried charged residues. These methods, however, are unable to detect the more subtle structural inconsistencies often localized in nonconserved loops, and cannot provide an assessment of the correctness of their geometry.

The accuracy of the model determines the extent to which it can be used. "Low-resolution" models, those derived from templates sharing less than 70% residue identity with the target, are helpful to rationalize site-directed mutagenesis experiments aimed at the identification of residue essential for a given molecular recognition and binding process. On the other hand, models based on more closely related templates may be useful

during a compound optimization process. For example, models of closely related species variants of an enzyme may allow the optimization of drug specificity.

We wish to thank all the users of Swiss-PdbViewer and SWISS-MODEL that have sent feedback, ideas, comments, and suggestions and thereby have helped to improve the workbench. Special thanks to Jeff Shaw for his constructive comments and patient β-testing of the successive PC versions.

Received June 10, 1997

6 References

[1] Bairoch, A., Apweiler, R., *Nucleic Acids Res.* 1996, *24*, 21–25.
[2] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M., *J. Mol. Biol.* 1977, *112*, 535–542.
[3] Peitsch, M. C., *Bio/Technology* 1995, *13*, 658–660.
[4] Peitsch, M. C., *Biochem. Soc. Trans.* 1996, *24*, 274–279.
[5] Peitsch, M. C., Wilkins, M. R., Tonella, L., Sanchez, J.-C., Appel, R. D., Hochstrasser, D. F., *Electrophoresis* 1997, *18*, 498–501.
[6] Chothia, C., Lesk, A. M., *EMBO J.* 1986, *5*, 823–826.
[7] Bajorath, J., Stenkamp, R., Aruffo, A., *Prot. Sci.* 1993, *2*, 1798–1810.
[8] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., *J. Mol. Biol.* 1990, *215*, 403–410.
[9] Pearson, W. R., Lipman, D. J., *Proc. Natl. Acad. Sci. USA* 1988, *85*, 2444–2448.
[10] Huang, X., Miller, M., *Adv. Appl. Math.* 1991, *12*, 337–357.
[11] Blundell, T., Sibanda, B. L., Sternberg, M. J., Thornton, J. M., *Nature* 1987, *326*, 347–352.
[12] Jones, D. T., Thirup, S., *EMBO J.* 1986, *5*, 819–822.
[13] Greer, J., *Proteins Struct. Funct. Genet.* 1990, *7*, 317–334.
[14] Ponder, J. W., Richards, F. M., *J. Mol. Biol.* 1987, *193*, 775–785.
[15] Lüthy, R., Bowie, J. U., Eisenberg, D., *Nature* 1992, *356*, 83–85.
[16] Sippl, M. J., *Proteins Struct. Funct. Genet.* 1993, *17*, 355–362.
[17] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., Karplus, M., *J. Comp. Chem.* 1983, *4*, 187–217.
[18] Peitsch, M. C., Herzyk, P., Wells, T. N. C., Hubbard, R. E., *Receptors Channels* 1996, *4*, 161–164.
[19] Herzyk, P., Hubbard, R. E., *Biophys J.* 1995, *69*, 2419–2442.
[20] Guex, N., Peitsch, M. C., *Protein Data Bank Quarterly Newsletter* 1996, *77*, 7.
[21] Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., Weiner, P. K., *J. Am. Chem. Soc.* 1984, *106*, 765–784.
[22] Morris, A. L., MacArthur, M. W., Hutchinson, E. G., Thornton, J. M., *Proteins Struct. Funct. Genet.* 1992, *12*, 345–364.
[23] IUPAC-IUB Commission on Biochemical Nomenclature, *J. Mol. Biol.* 1970, *52*, 1–17.
[24] Appel, R. D., Bairoch, A., Hochstrasser, D. F., *Trends Biochem. Sci.* 1994, *19*, 258–260.
[25] Birktoft, J. J., Rhodes, G., Banaszak, L. J., *Biochemistry* 1989, *28*, 6065–6081.
[26] Kelly, C. A., Nishiyama, M., Ohnishi, Y., Beppu, T., Birktoft, J. J., *Biochemistry* 1993, *32*, 3913–3922.
[27] Wilks, H. M., Hart, K. W., Feeney, R., Dunn, C. R., Muirhead, H., Chia, W. N., Barstow, D. A., Atkinson, T., Clarke, A. R., Holbrook, J. J., *Science* 1988, *242*, 1541–1544.
[28] Birktoft, J. J., Carnhagan, G. E,, Rhodes, G., Roderick, S. L., Banaszak, L. J., *Biochem. Soc. Trans.* 1989, *17*, 301–304.
[29] Guex, N., Widmer, F., Gaillard, P., Carrupt, P.-A., Testa, B., in: *Trends in QSAR Mol. modelling* 1993, *9*, 485–486.
[30] Gleason, W. B., Fu, Z., Birktoft, J. J., Banaszak, L., *Biochemistry* 1994, *33*, 2078–2088.
[31] White, J. L., Hackert, M. L., Buehner, M., Adams, M. K., Ford, G. C., Lentz, P. J., Smiley, I. E., Steindel, S. J., Rossmann, M. G., *J. Mol. Biol.* 1976, *102*, 759–779.
[32] Tilton, R. F., Dewan, J. C., Petsko, G. A., *Biochemistry* 1992, *31*, 2469–2481.
[33] Harrison, R. W., Chatterjee, D., Weber, I. T., *Proteins Struct. Func. Genet.* 1995, *23*, 463–471.