Protein Fold Recognition and Comparative Modelling using HOMSTRAD, JOY and FUGUE

Ricardo Núñez Miguel, Jiye Shi and Kenji Mizuguchi

Department of Biochemistry, University of Cambridge
80 Tennis Court Road, Cambridge CB2 1GA, UK

kenji@cryst.bioc.cam.ac.uk

in

Protein Structure Prediction: Bioinformatic Approach / edited by Igor F. Tsigelny
ISBN 0-9636817-7-X
La Jolla: International University Line; 2002. pp. 143-169.
www.iul-press.us

Abstract

This article illustrates how tools exploiting the knowledge of protein three-dimensional structure can be used to identify homologues of known structure, generate sequence-structure alignments and assist model building. The tools described here include HOMSTRAD, a database of structure-based alignments for protein families of known structure, JOY, a program to annotate local environments in structure-based alignments, and FUGUE, a program to perform sequence-structure homology recognition. After a brief review of the whole process of homology recognition and comparative modeling, a specific example clarifies all the steps involved. This type of analysis will help obtain a better understanding of the function of many proteins whose sequences are known.

INTRODUCTION

Divergent evolution has given rise to families of homologous proteins, where members of a family share similar but often diverged amino acid sequences. Even though these distantly related members have little sequence similarity, their three-dimensional (3D) structures are very well conserved and they also share, broadly speaking, common functions. Thus, if we can somehow assign an unknown protein sequence to a known family, which has a member of known structure, we can learn about the structure and function of this unknown protein (Fig. 1). This is the basis of structure prediction and functional inference using sequence-structure homology recognition. This type of analysis can bridge two traditional branches of biology, sequence database searches and structural studies. With the total number of complete genomes soon to exceed 200, and a growing number of experimentally defined 3D structures, it has huge potential for providing a new type of knowledge in the post-genomic era.

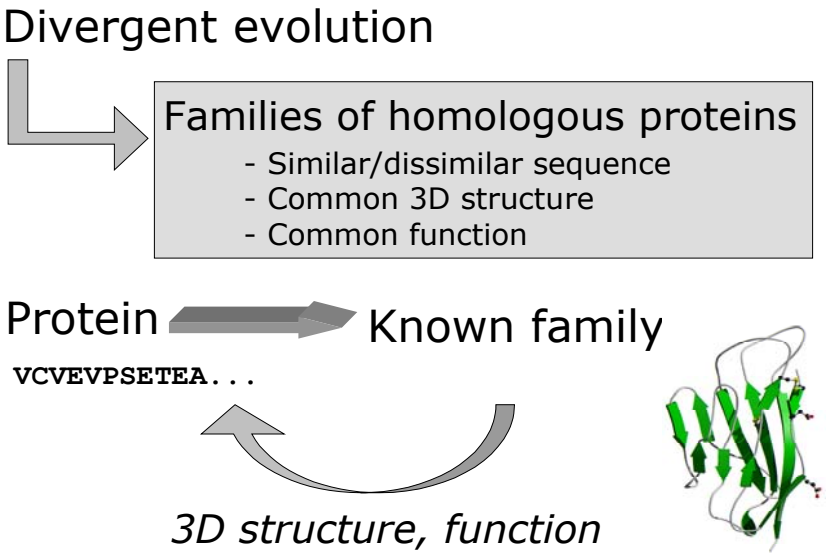

Fig. 1. From divergent evolution to 3D protein structure and function.

We have developed various tools to facilitate many of the important steps in structure/function prediction using homology recognition. The database HOMSTRAD[1] (http://www-cryst.bioc.cam.ac.uk/homstrad/) provides information about protein families with known structure and presents a curated collection of structure-based

alignments of the members of these families. We take all known protein structures, cluster them into families and align the sequences of the representative members of each family on the basis of their structures. The alignments are generated by the program COMPARER[2] and several other tools to optimize the conservation of local environments and are individually checked. Because it provides structural alignments, it can be used to evaluate sequence alignments, as a standard benchmark set[3] or by direct comparison with sequence alignments in Pfam.[4] Because it is manually curated, it can even be used to benchmark automatic structure comparison methods.[5]

We use HOMSTRAD in many ways, but perhaps the most important application is the derivation of environment-specific substitution tables.[6] Each residue in a protein structure stays in a particular local environment, which can dramatically influence the amino acid substitution pattern of the residue. For instance, it is well known that residues buried in the core of the structure are more conserved than those on the surface, and residues within secondary structure elements (SSEs) are more conserved than those in coil regions. One example is illustrated in Fig. 2, which shows probabilities that a residue will not be substituted by any other residue type during evolution. These probabilities were calculated from the structure-based alignments in the HOMSTRAD database. Not only does a buried position have a higher conservation probability than a surface position for all 20 amino

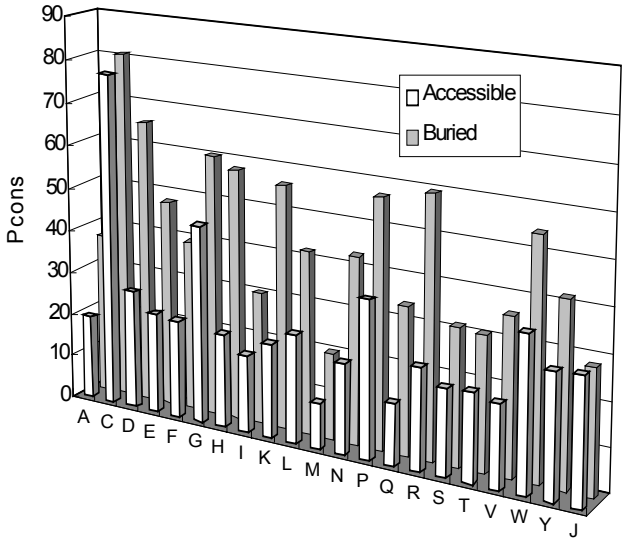

Fig. 2. Probabilities that a particular amino acid residue will not be substituted by any other residue type during evolution. The data were calculated from selected structure-based alignments in the HOMSTRAD database. Disulphide-bonded cysteine (C) and non-disulphide-bonded cysteine (J) residues are distinguished.

acids, the figure also shows that the increases in conservation from surface to buried residues are not uniform, i.e., the residues that undergo the largest increases in conservation are polar or charged, a typical example being asparate (represented as D). This indicates that local environments contain useful information for predicting amino acid substitution patterns.

The program JOY[7] (http://www–cryst.bioc.cam.ac.uk/~joy/) can define these local environments and annotate structure-based alignments. It produces formatted alignments, in which normal one-letter amino acid codes are decorated with special symbols to allow easy identification of local environments (see Fig. 8). The program has proved to be a useful tool in examining and optimizing sequence-structure or structure-structure alignments and identifying distant homologues.[8,9] A JOY-formatted alignment highlights unique patterns of amino acid substitutions in various environments. For example, the conservation of buried asparate residues can be easily recognised, as these are shown in bold capital letters. Thus, it helps to identify misaligned regions or residues that play important structural roles.

Using JOY and structure-based alignments in HOMSTRAD,[1] we have derived amino acid substitution matrices for different environments.[6] These are one of the essential elements of the homology recognition program FUGUE[10].

FUGUE is a tool developed to associate a query sequence with its homologues of known structure. It compares a query sequence or sequence alignment against each structural profile in its profile library derived from HOMSTRAD and assesses the compatibility between the sequences and the structures. A structural profile consists of two matrices: a scoring matrix and a gap penalty matrix. The key feature of FUGUE is to calculate both matrices not only according to the amino acid sequence information, which is used by traditional sequence-only fold recognition methods, but also the local structural environment information.

Traditional sequence-only methods ask the question “what is the likelihood of amino acid A being substituted by amino acid B during evolution?” In contrast when

we construct the scoring matrix for the FUGUE structural profile, we ask the question "what is the likelihood of amino acid A, within structural environment E, being substituted by amino acid B during evolution". FUGUE uses 64 environments defined by the combination of three structural features: main-chain conformation and secondary structure (helix/strand/coil/positive phi torsion angle), solvent accessibility (accessible/inaccessible) and hydrogen bonding status (true or false for: side-chain to main-chain NH/side-chain to main-chain CO/side-chain to other side-chain). The local environment is calculated for each residue of the structure and the corresponding environment-specific amino acid substitution pattern is stored in the scoring matrix of the structural profile.

During divergent evolution, insertion/deletions occur more frequently on the surface region of the protein than in the core region and also more frequently within the coil region than within SSEs. In sequence alignments, insertions/deletions are represented as gaps. FUGUE calculates the gap penalty matrix according to the local structure information. For instance, positions in SSEs and core regions receive higher gap penalties than those in coil and surface regions and positions at the center of an SSE receive higher gap penalties than those at the terminal of an SSE. These structure-dependent gap penalties are the second essential element of FUGUE.

In this article, we illustrate how these and other tools can be used to identify homologues of known structure, generate sequence-structure alignments and assist model building. After briefly reviewing the whole process of fold recognition and comparative modeling, we first describe some practical considerations in using FUGUE, which plays a key role in the whole process. We then use a specific example to illustrate all these steps, including discussions on various other tools.

OVERVIEW

Our goal is to assign an unknown sequence to a family with known structures and build an accurate model for the 3D structure of the protein, which then will allow functional inferences. There are several steps to achieve this goal. First, given a target protein sequence, one or more homologous proteins of known structure need to be

identified. Second, it is important to have a good sequence alignment between the homologues and the target protein. These two steps are crucial; if the proteins identified are not true homologues, or if the two amino acid sequences are wrongly aligned, the 3D-model obtained will be wrong even if the rest of the process is perfect. In the next two sections we will discuss how FUGUE can play crucial roles in these two steps.

The next step in the process consists of obtaining the structure from the alignment by using one of the available comparative modeling programs, followed by the refinement of the obtained structure. Finally, a check of the structure is needed to avoid implausible models. If the protein structure includes some unlikely or impossible features, we go back to the alignment and try to improve it. If the alignment cannot be improved or alternative alignments do not lead to better models, it is possible that the selected homologues might be incorrect and new homologous proteins should be identified. Fig. 3 shows this process schematically.

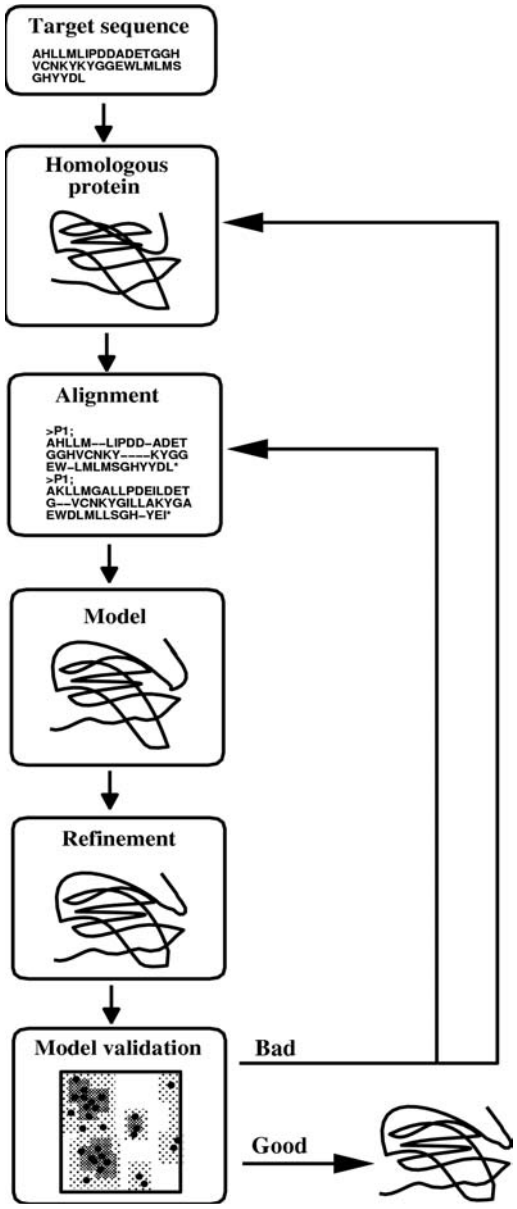

Fig. 3. Schematic representation of the steps followed in comparative modeling.

For each step described above, there are several good tools that can be used and in some cases it is a good idea to use more than one tool and select the best results.

IDENTIFICATION OF HOMOLOGUES

FUGUE is available to the public via a web server at http://www-cryst.bioc.cam.ac.uk/fugue. Given a single query sequence, the server runs PSI-

BLAST[11] to perform a search against the NCBI non-redundant sequence database and collects sequence homologues. The alignment produced by PSI-BLAST is then used to calculate a sequence profile, which describes the observed amino acid distribution at each position of the query sequence. FUGUE compares this sequence profile against each structural profile in its library derived from HOMSTRAD and assesses the compatibility between the sequences and the structures.

The sequence homologues retrieved by PSI-BLAST provide valuable information about the sequence family, which can improve the performance of FUGUE. However, in some cases non-homologous sequences (false positives) may be included in the PSI-BLAST alignment and the alignment itself may contain serious errors. Advanced users are recommended to check the PSI-BLAST alignment when receiving the FUGUE results. They can improve the alignment and re-submit it to the FUGUE server by selecting the option that tells the server to use the input alignment for sequence-structure comparison. There is also an option on the FUGUE server to skip the PSI-BLAST search and use a single input sequence for the search. This option should only be used when the user fails to obtain an alignment of reasonable quality between the query sequence and its sequence homologues.

During the database search, FUGUE aligns the query sequence profile against each structural profile using the scoring and gap penalty matrices stored in it. The query sequence profile is then randomized by 100 times and an alignment score is calculated for each randomized profile. A Z-score is calculated by comparing the alignment score for the original sequence profile against the scores for the randomized ones. Higher Z-scores indicate better compatibility between the query sequence and the structure and greater probability of homology.

FUGUE was benchmarked using a test set developed by Lindahl and Elofsson.[12] In the test set, 976 proteins of known structure are clustered into families, superfamilies and folds based on the SCOP[13] classification. An all-against-all recognition test can be carried out to check how well the program being benchmarked can re-establish the correct relationships among those proteins. Figure 4 shows the

benchmark result for FUGUE in recognizing protein pairs that share family level similarity, together with the results for some other fold/homology recognition tools provided by Dr. Elofsson. FUGUE significantly outperformed other methods. For example, at 99% specificity (i.e. 1 error out of every 100 predictions of homology), FUGUE obtained a sensitivity of 49% (i.e. 49% of true homologous protein pairs were recognized), while the best performance of other methods, obtained by HMMER-PSIBLAST, hit 42% sensitivity. Z-score confidence thresholds were estimated from the benchmark result. Specificities of 99% and 95% corresponded to Z-scores of 5.6 and 4.6, respectively. In practice, we set the default Z-score thresholds at 6.0 for 99% confidence and 5.0 for 95%.

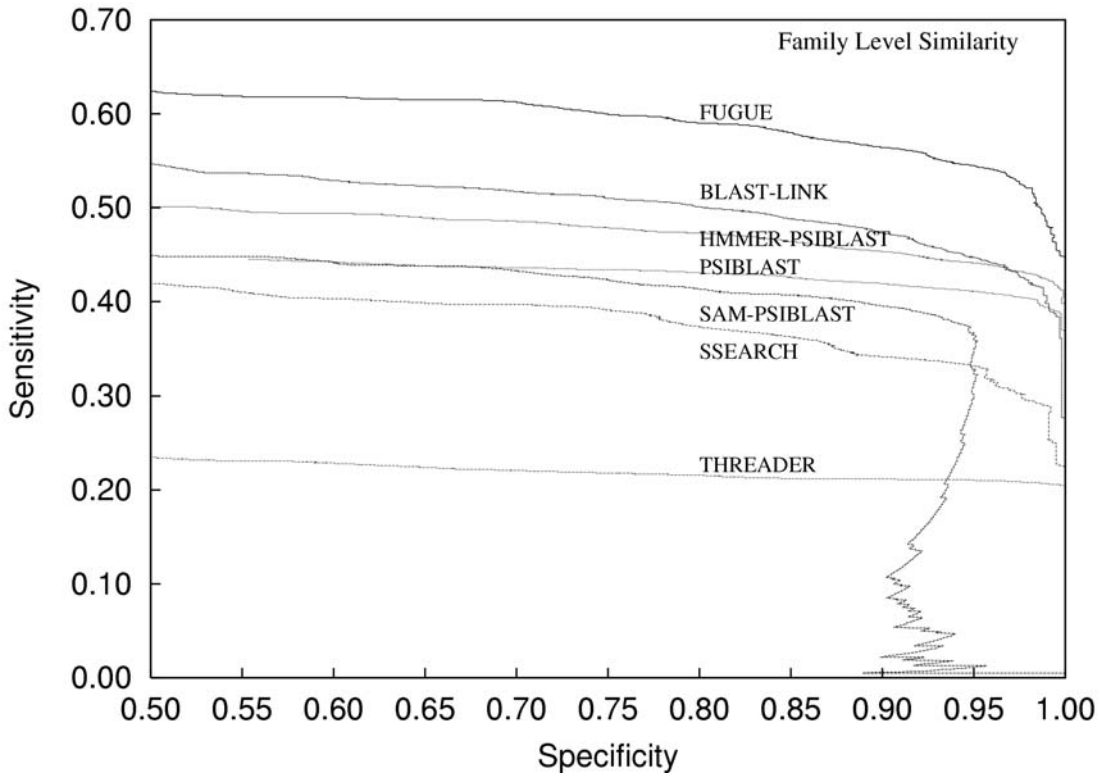

Fig. 4 Specificity-sensitivity curves of recognition performance at the family level using the test set provided by Dr. Elofsson.[12] Data other than that of FUGUE were kindly provided by Dr. Elofsson.

The recognition performance of FUGUE has also been benchmarked in two independent assessment exercises: CAFASP2 (http://cafasp.bioinfo.pl/) and LiveBench2 (http://bioinfo.pl/LiveBench/). FUGUE was ranked among the top servers, and was also key to the success of the Blundell group in CASP4.[14] This demonstrated

the usefulness of environment-specific substitution scores and structure-dependent gap penalties in homology recognition.

GENERATING SEQUENCE-STRUCTURE ALIGNMENT

FUGUE can also be used as a sequence-structure alignment program. The environment-specific substitution scores and structure-dependent gap penalties can also help to build more accurate sequence-structure alignments compared with many sequence-only alignment programs like CLUSTALW,[15] especially when the percentage sequence identity (PID) is low and the structural information becomes more significant. By using Fischer's benchmark test-set,[16] we observed that FUGUE outperforms both CLUSTALW and GenTHREADER[17] in alignment accuracy.[10]

The FUGUE homology recognition server searches for homologues in the structural profile library and automatically generates the best alignments for the top hits. This step can be shortened, however, if some homologues of known structure are already known. For example, suppose we are interested in a protein, which is known to belong to the aspartic proteinase family. Rather than submitting this sequence to the FUGUE homology recognition server, we can directly go the asparatic proteinase page of HOMSTRAD (http://www-cryst.bioc.cam.ac.uk/cgi-bin/homstrad.cgi?family=asp), This page can be reached using the search facility, either with a keyword (type in 'aspartic'), or the PDB code of a homologue if it is known (type in '5pep'). A quick BLAST search is also available (http://www-cryst.bioc.cam.ac.uk/cgi-bin/homstrad/blast.cgi). Once the aspartic proteinase page has been located, the user can simply click the blue 'ALIGN' icon at the top left corner. This will allow the submission of a user's own sequence and FUGUE will generate the optimal sequence-structure alignment.

EXAMPLE

All the steps described above for homology recognition and comparative modeling will be illustrated using a particular example. NDP-4-keto-6-deoxyglucose 3,5-epimerase[18] (EvsA) from *Amycolatopsis orientalis* is involved in the production of NDP-4-*epi*-vancosamine, an L-amino-2,6-dideoxysugar needed in the biosynthesis of

the heptapeptide antibiotic chloroeremomycin,[19] which is effective in combatting infections from *Staphylococcus aureus.* EvsA epimerizes the positions three and five of the sugar ring (Fig. 5). Two molecules of NDP-4-*epi*-vancosamine are attached to the antibiotic heptapeptide backbone.

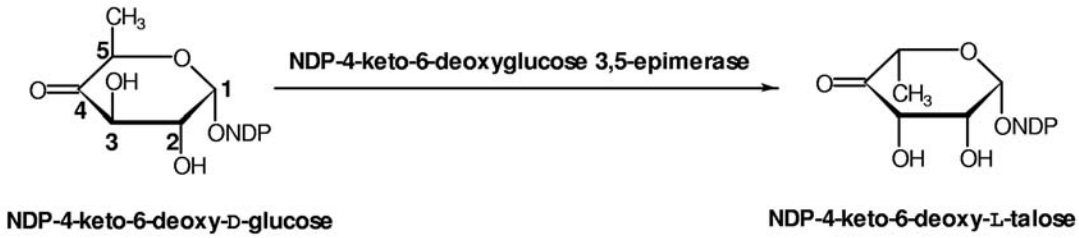

Fig. 5. Chemical reaction catalyzed by EvsA from *Amycolatopsis orientalis*.

Searching for homologues

The amino acid sequence of EvsA can be obtained from SWISSPROT/TrEMBL (accesion code: O52806) and is 205 residues long. We can easily check if there are close homologues of known structure, using a BLAST[11] search against the PDB. There are numerous web servers such as the one at NCBI (`http://www.ncbi.nlm.nih.gov/blast/index.html`). For EvsA the BLAST search detected two statistically significant hits. Both are the same enzyme dTDP-6-Deoxy-D-Xylo-4-Hexulose 3,5-Epimerase (RmlC) from different organisms. The first hit is RmlC from *Salmonella typhimurium* (PDB code 1DZR Chain A) with an E-value of $1 \cdot 10^{-22}$, a PID of 33%, an alignment of 175 residue long and with 0% gaps. The second hit is RmlC from *Methanobacterium thermoautotrophicum* (PDB code 1EP0, Chain A) with an E-value of $1 \cdot 10^{-23}$, a PID of 33%, an alignment of 167 residue long and with 2% gaps. The structures of these proteins consist entirely of β strands and the fold is called a double-stranded β-helix. Each turn of the helix is made up of two pairs of anti-parallel strands that are linked with short turns. RmlC catalyzes the same reaction as EvsA (Fig.5), the only difference being the nucleotide moiety of the substrates. It is deoxythymidine diphosphate (dTDP) in the case of RmlC whereas it is unclear in the case of EvsA.

It is not always possible to find close homologues of known structure with high PIDs. If BLAST does not detect any homologous protein in the PDB, it is necessary to perform additional analyses. Even if there are close homologues, as in our present example, it is always a good idea to perform these additional analyses, as they provide more information that can assist the alignment and model building processes.

A first, and probably the most useful, piece of information can be derived from a homology search against a bigger sequence database, for example, the NCBI non-redundant database. This can be carried out, again, with the BLAST or PSI-BLAST programs, either running the program locally or using a web server. An advantage of running the program locally is that we can process the output and use various other tools. For example, the BLAST alignment can be converted into a FASTA formatted file using the program blastalign2fasta in the SEALS package.[20] This alignment can be viewed and examined, or sent directly to other programs such as FUGUE.

In the current example of EvsA, our BLAST search against the non-redundant database detected 106 homologues. The closest are the EvsA proteins from: *Streptomyces griseus, Streptomyces peucetius, Streptomyces galilaeus, Streptomyces overmitilis, Streptomyces nogalater and Saccharopolyspora erythraea*. The alignment revealed the following conserved residues: D21, R23, G24, Q48, S51, V58, R60, G61, H63, K73, V75, G80, D84, D88, R90, S93, W99, H120, F122, Y133, Y139, D151, S167 and D170. Even though no direct structural information is available for any of these homologues, these conserved residues are likely to play important roles, by either stabilizing the structure, or being involved in catalysis.

A second piece of information, universally available, is the secondary structure prediction of the target protein. There are several programs that predict the secondary structure, for example: PSI-PRED[21] (http://insulin.brunel.ac.uk/psipred/), PHD[22] (http://www.embl-heidelberg.de/predictprotein/predictprotein.html), SSPRED2[3] and PREDATOR.[24] In the current example, the JPRED[25] server (http://jura.ebi.ac.uk:8888/) was used. The consensus JPRED prediction indicates three helices at sequence positions 31-38, 176-181 and 188-196, and 10 strands at positions: 11-15, 26-28, 45-49,

58-64, 73-78, 82-89, 98-104, 110-115, 122-125 and 129-135. Many secondary structure prediction programs, in fact, use the alignment obtained from a BLAST search, thus, these two analyses can be combined and performed at once.

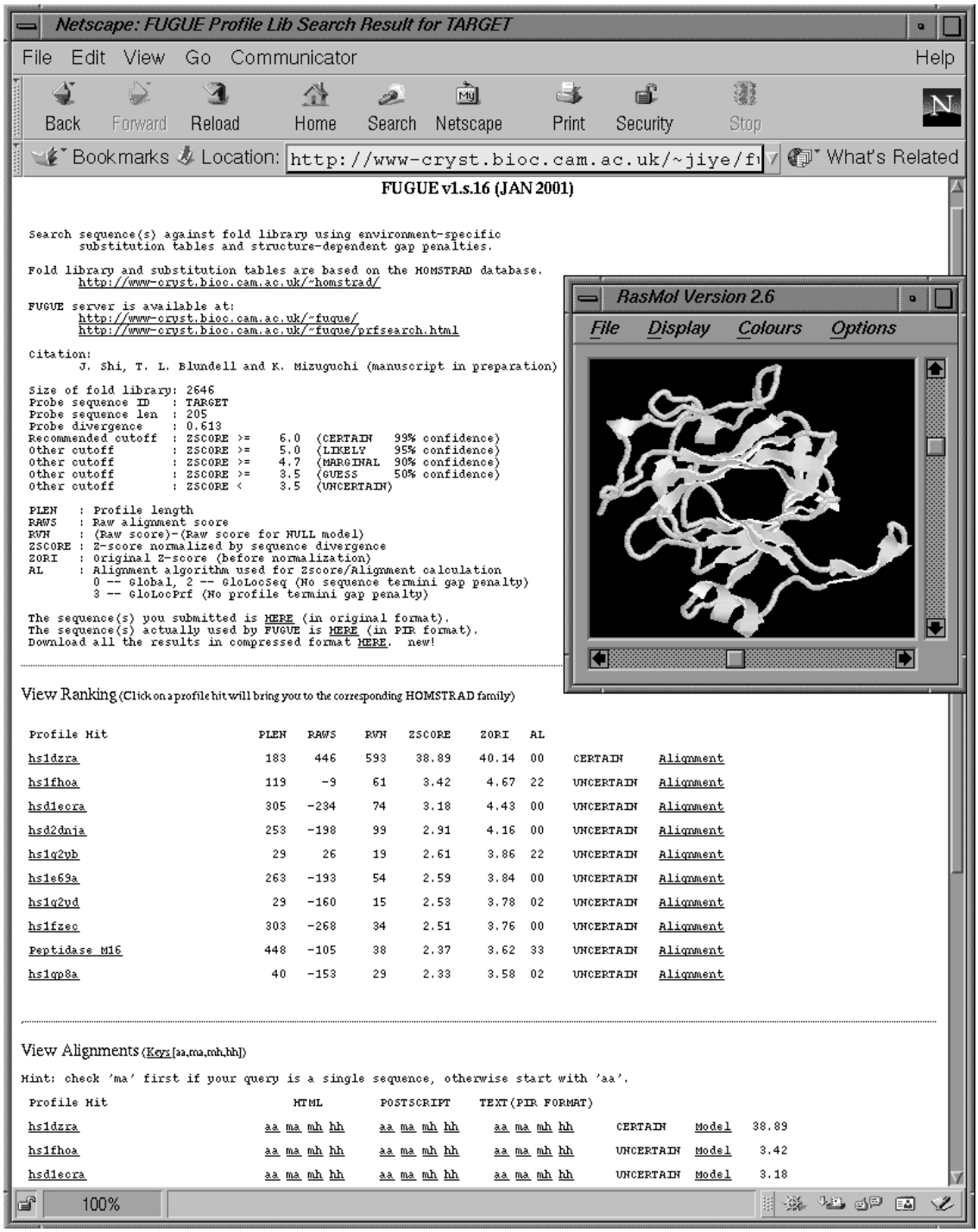

Fig. 6. Prediction results from the FUGUE server using the sequence of EvsA as query.

These two analyses, sequence database searching and secondary structure prediction can be effectively combined with more sophisticated fold or homology recognition programs (see URL for LiveBench, for examples). These programs can recognize homologues of known structure, which may not be detected by BLAST or other sequence-only methods. We used FUGUE to search for homologues of EvsA.

Fig. 6 shows the prediction result from the FUGUE server using the sequence of EvsA as a query. The output consists of three sections: header, rank and alignment. Here we explain the first two sections in detail and the alignment section will be explained later when we discuss sequence-structure alignment.

The header section gives the version number of FUGUE, the size of the current HOMSTRAD structural profile library, the divergence of homologous sequences collected by PSI-BLAST, the confidence level of different Z-score values and the explanations of abbreviations used in the rank section. Note that the sequence divergence is used by FUGUE internally to adjust Z-score values and should normally be ignored by the user.

The rank section lists the top 10 hits found by FUGUE, ranked by Z-score. The first column gives links to the corresponding HOMSTRAD family pages of the hits, where more structural/functional information can be obtained. The fifth column is the Z-score. The eighth column translates the Z-score into one of the five more understandable categories of prediction assessment: certain, likely, marginal, guess and uncertain, with decreasing confidence levels.

In our example, FUGUE searched 2646 HOMSTRAD families and predicted that the family hs1dzra (RmlC from *Salmonella typhimurium*, PDB-ID: 1dzr) is the most compatible structure of EvsA with a significant Z-score of 38.89, which corresponds to the confidence level of “certain”. The second hit has a Z-score of 3.42, which indicates an “uncertain” prediction. Thus, according to the FUGUE result, we have good confidence on the prediction that the first hit is a homologue of EvsA and it can be used as a structural template for EvsA in comparative modeling**.**

Alignment

The next step, and one of the most important, is to obtain a good alignment between the template protein(s) and the target protein(s). The FUGUE server produces alignments between the query sequence and entries of the HOMSTRD databases and annotates the alignments using JOY[7]. The annotated alignments are useful for examining whether particular local environments are compatible with the query sequence and its homologues.

The alignment section in the FUGUE output (see Fig. 6) gives alignments between the query sequence and each of the top 10 hits in three formats: HTML for online browsing, PostScript for printing and pure text format for using with other programs. The HTML and PostScript versions are produced by JOY for annotations of the structure. Four types of alignments are available. The "aa" type is the alignment between the query sequence, together with its homologues collected by PSI-BLAST (all sequences), and all the structures of the corresponding HOMSTRAD family (all structures). This is the most informative type, as all the available sequence and structure information is shown. However, when there are a large number of sequence homologues collected by PSI-BLAST, this type of alignment is difficult to examine by eye. In such situations, the "ma" type, which removes the sequence homologues from the "aa" type (showing only the master sequence), can be used for visual inspection. The "mh" type is the alignment between the query sequence and the single structure, which has the highest PID to the query, in the HOMSTRAD family (master against the structure with the highest PID). It can be directly used as the input for comparative modeling software. The "hh" type represents the most similar sequence-structure pair, in terms of PID, in the "aa" type alignment.

FUGUE also builds rough models (see the "model" column in the alignment section) for the query sequence by using "mh" type alignments of the top 10 hits. For each model, backbone coordinates are copied from the template structure according to the sequence-structure alignment. For the residues in the query sequence, which do not have corresponding residues in the template structure, no coordinates are predicted. The

rough models can help the user to assess the confidence of homology recognition and the alignment quality. For example, a model consisting of only short fragments suggests either a poor alignment or a non-homologous sequence-structure pair.

In Fig. 6, a rough model built by using the "mh" type alignment of the top hit is shown in Rasmol[26] in the upper-right corner. The model maintains most of the basic structural elements of the template structure, which, together with a significant Z-score, suggests that the FUGUE alignment between EvsA and the HOMSTRAD family hs1dzra be of reasonably good quality.

It is important to confirm that most of the conserved residues in the family alignment of the target are correctly aligned with the template. In our example all conserved residues, previously mentioned, are also conserved in the template, indicating that the FUGUE alignment is reliable. It is also important to check if the active site residues in the template protein(s) are conserved in the target. In our example the active site residues, reported in the literature[27], in RmlC are: Phe20, Arg24, Phe27, Glu29, Gln48, Asn50, Arg60 and Tyr139. All these residues are conserved in EvsA except for Phe27, Glu29 and Asn50. These three residues in RmlC are involved in binding the nucleotide, suggesting that the nucleotide moiety of the substrate for EvsA may be different.

Another way of checking the conservation of functional residues is to go back to the template entry in HOMSTRAD. Clusters of conserved residues are often observed in the amino acid sequences of proteins with a common function. Such conserved clusters, usually called patterns, motifs or fingerprints, are catalogued in databases such as PROSITE,[28] BLOCKS,[29] PRINTS,[30] and PROF_PAT.[31] HOMSTRAD[4] has incorporated 644 PROSITE patterns, as decorations in the family alignments. Fig. 7 shows an example of a HOMSTRAD family "Muconate lactonizing enzyme-like" that has two PROSITE patterns, one of which is shown in the figure. The occurrence of the family pattern(s) in the query sequence is often a good indication that the homologues have been chosen correctly and that the alignment is reasonable.

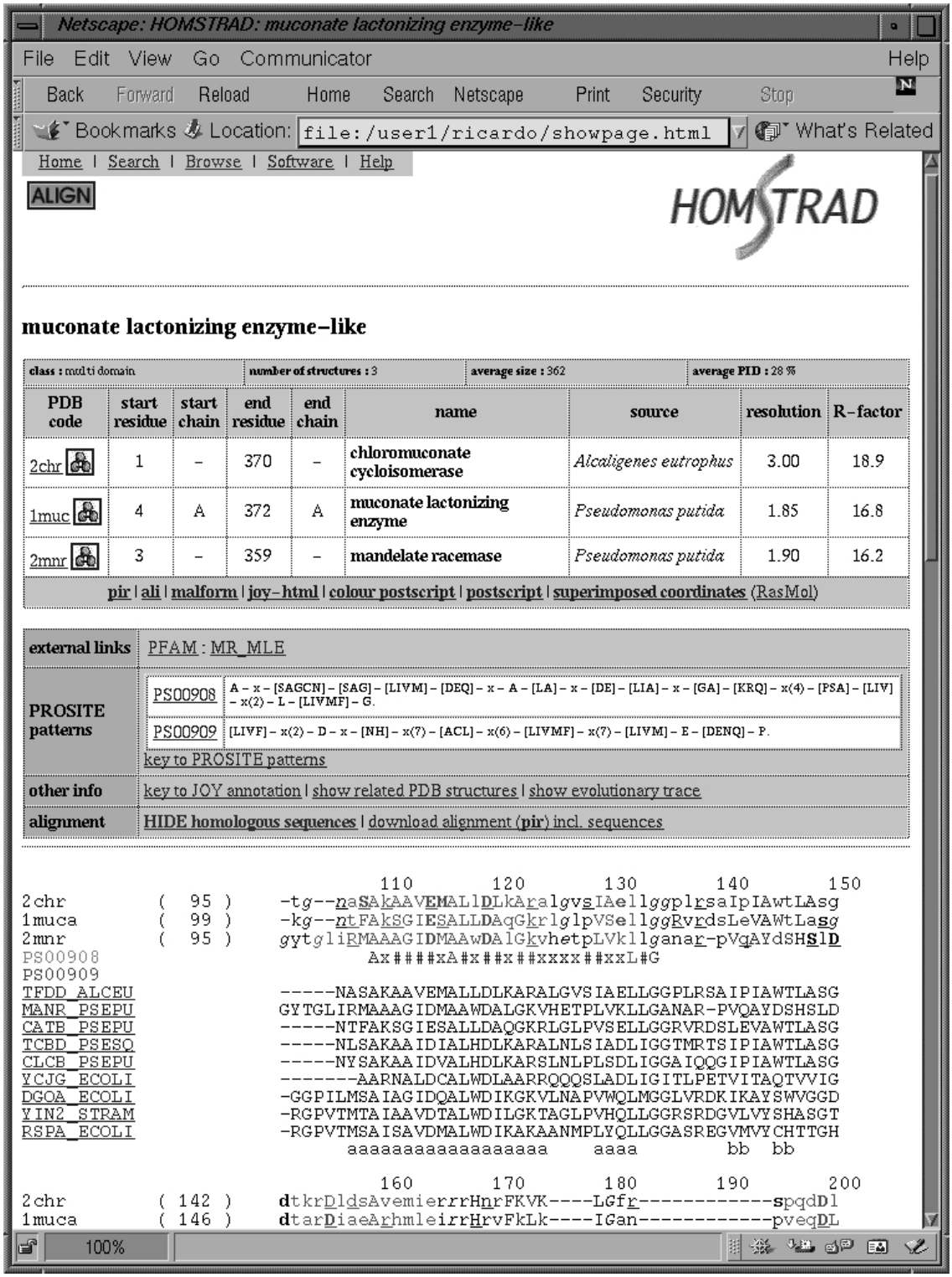

Fig. 7. Muconate lactonizing enzyme-like HOMSTRAD family aligned with two PROSITE motives and showing one of them.

Sequence alignment algorithms are dependent on adjustable parameters whose values determine the sequence similarity, placement of gaps and the ultimate alignment. Even

using the best alignment program, manual adjustments may be necessary. This is partly because most programs do not utilize all available information. For example, FUGUE uses a PSI-BLAST alignment as input, thus takes into account the conserved residues of the target protein. It does not, however, use any functional information such as the positions of active site residues. In the present version of FUGUE, information about predicted secondary structures is not used. Therefore, it is possible to improve the

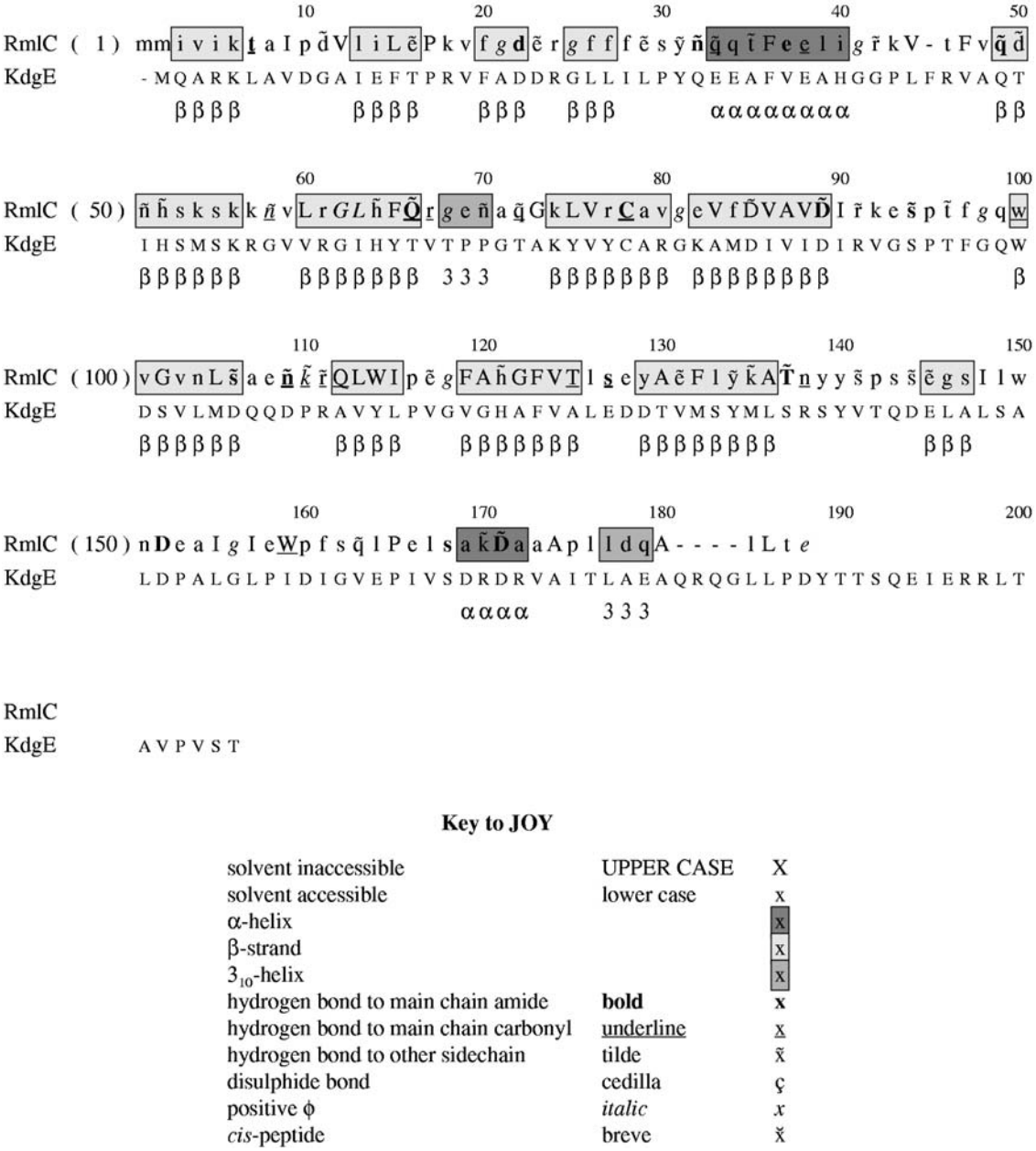

Fig. 8. Sequence alignment of target EvsA with parent RmlC and formatted by JOY.

alignment by altering active site residues or by maximizing the consistency between predicted secondary structures and those in the template structures. Fig. 8 shows the FUGUE alignment after a small manual modification, in which the last four residues in RmlC have been moved four position to the right in order to align the two leucines in both proteins and the glutamic acid to the aspartic acid.

Modeling

Once a good alignment is available, a 3D-structural model of the target protein can be built. In order to obtain the model several programs can be used, including COMPOSER[32] and MODELLER[33]. These two programs produce an output file in PDB format containing the 3D coordinates of every non-hydrogen atom including all loops, the N- and C-termini and side chains. Other programs, such as SCORE[34], produce only the backbone of the conserved parts of the protein. After the backbone, the structurally variable regions, which normally corresponded to the loops, can be obtained using Sloop,[35] CODA[36] or the Loop Database of SYBYL.[37] CODA runs two programs for the prediction of the structurally variable regions of protein structures: FREAD, a knowledge-based method using a database of fragments taken from the PDB and PETRA, an *ab initio* method using a database of computer generated conformers. CODA is available on the web at "http://www-cryst.bioc.cam.ac.uk/~charlotte/Coda/search_coda.html". CODA is helpful for solving the problem generated by the insertions where there is no template for these residues.

When the backbone is available the side chains may be added. To do this, programs such as SCRWL[38] or CELIAN[39] can be used. The replacement of side chain residues often results in unfavorable interactions such as steric overlaps between atoms. Relaxing these bad side chain contacts requires repositioning side chain atoms while fixing the backbone, to seek a local energy minimum. When that procedure cannot relax a local conformation from a high energy state, the original alignment may require adjustment.

Heteroatoms

In some cases, it is important to obtain protein-cofactor, protein-substrate or protein-cofactor-substrate complexes, especially if the aim of the work is to study the reaction mechanism or for drug design. This is difficult, however, if the template structure does not include the cofactor or substrate. Sometimes in the PDB there are multiple entries of the same protein, with and without the coordinates of the cofactors and/or substrate, and a relevant template should be chosen depending on the purpose of the study. The program MODELLER allows for the building of cofactors/substrates. The program SCORE, however, does not build cofactors/substrates automatically. In this case a coordinate superposition between the model (without cofactors) and the template (with cofactors) can be obtained using programs such as MNYFIT[40] and the coordinates for the cofactors/substrates can be transferred.

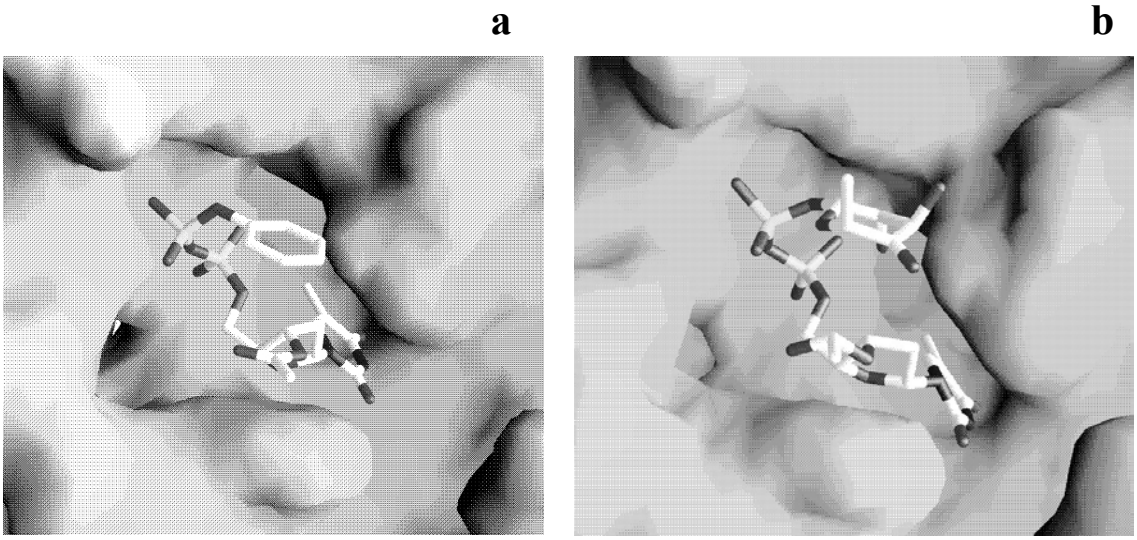

Fig. 9. a) Initial substrate analog of RmlC from *Salmonella typhimurium* placed in the EvsA model. b) Modifyed substrate for EvsA. Figures generated by GRASP.[48]

The PDB entry 1DZT is a structure of RmlC complexed with the substrate analog 3'-O-acetylthymidine-5'-diphosphate-phenyl ester. Using this entry as the template, we have built, using MODELLER, the model of EvsA complexed with this molecule (Fig. 9a). Even though the experimental evidence to define the nucleotide moiety of the substrate is inconclusive,[41] we assume that 3'-O-acetylthymidine-5'-diphosphate-phenol is a substrate analog for EvsA. This substrate analog can be modified using programs such as: SYBYL and InsightII[42] to obtain the coordinates of

the real substrate. Fig. 9b shows the result after the modification of the substrate analog with InsightII. The phenyl group has been transformed to the sugar molecule.

Refinements

After generating initial backbone and side chain conformations, the entire structure is energy minimized. Since energy minimization only finds a nearby local minimum for a given initial structure, molecular dynamic and Monte Carlo techniques are sometimes used to seek more energetically favorable structures.

There are several methods to evaluate what parts of the model are not well modeled. Apart from visual inspection, PROCHECK,[43] Verify3D[44] or PROSAII[45] can be used. After the energy minimization some residues can still present wrong torsion angles in the Ramachandran plot or negative values in the Verify3D output. In these cases, the templates should be checked to see whether the problem has been carried over from the template structure. If the problematic residues are placed in a loop, the entire loop can be remodelled, with CODA, for example, selecting different loops until all test results become satisfactory.

Model validation

There are several methods for validating models. Programs such as Verify3D or PROSAII check whether the structure is reasonable from the perspective of sequence-structure compatibility and other programs such as PROCHECK examine the backbone and side chain stereochemistry.

Another good way to test the generated model is by realigning the model to their template structure using the structure alignment program COMPARER,[2] and annotate the alignment with JOY. This allows visual inspection of the conservation of both residues and their structural environments. A model can be validated by docking known substrates or inhibitors to the active site, if this information is known. The results should be consistent with the known specificity for substrates and inhibitors.

Finally, experimental evaluation can be performed. Site-directed mutagenesis on the active-site or binding regions can be carried out. Structurally important residues can be mutated to check if the mutation affects the folding of the protein.

Model

On the basis of these model validation methods, the model is either accepted, rejected, or the alignment is modified and the comparative modeling process repeated. Once the model is accepted, a large amount of useful information can be derived.

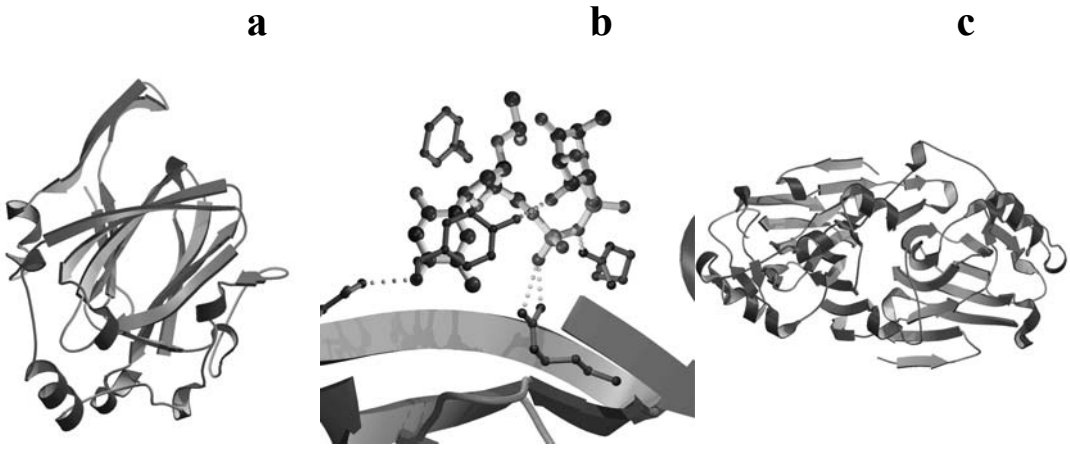

Fig. 10. a) Cartoon representation of the model of EvsA as a monomer. b) Representation of the active site of EvsA. Protein side chains (dark bonds) and substrate (light bonds) are illustrated. Hydrogen bonds are represented as dashed lines. c) Cartoon representation of the model of EvsA as a dimer.

Fig. 10 shows cartoon representations of the model obtained for EvsA by following the previously explained steps (drawn by the programs MOLSCRIPT[46] and RASTER3D[47]). Although RmlC is known to be a dimer, it is not known whether this is also true for EvsA. We have examined amino acid sequences and structural features in the putative dimeric interface of the model of EvsA and concluded that it is likely to be a dimer. A dimer model has been obtained for EvsA, Fig. 10c, in which the same dimeric interactions as in RmlC have been found. In addition residues from both subunits are involved in the active site.

One of the most important pieces of information provided by a model is the knowledge of the active site. The residues that bind the cofactor and/or the substrate by

hydrogen bond or salt bridge can be examined with JOY. For EvsA, the residues hydrogen-bonded to the substrate are: Gln A48, Arg A60, Tyr A139 and Arg B23. The last residue belongs to the second chain of the dimeric structure. In Fig. 10b all these residues as well as those involved in van der Waals contacts with the substrate are presented. Fig. 11 shows a schematic representation of the interactions of the substrate with the active site residues.

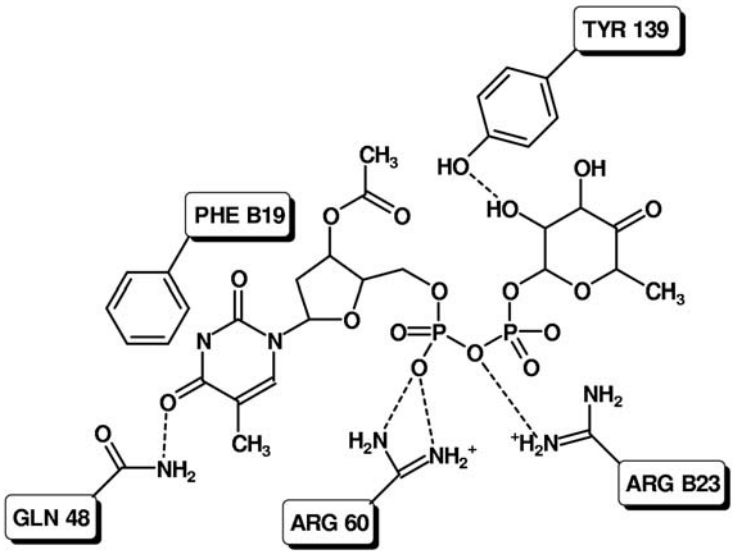

Fig. 11. Schematic representation of the interactions at the EvsA active site. Hydrogen bonds are represented as dashed lines.

In this example, the main aim for modeling the EvsA enzyme is to study its active site and find some possible mutation targets that could alter the enzymatic reaction. By altering one step of the biosynthesis of the antibiotic chloroeremomycin, *Amycolatopsis orientalis* may produce a new antibiotic, which cannot be recognized by antibiotic resistant bacteria. Fig. 12 shows part of the active site of EvsA (the figure has been produced by Rasmol[26]). Tyr 139 is hydrogen bonded to the hydroxyl group in the C2 of the sugar, while the hydroxyl group in the C3 has a closer residue, Thr 141, but the distance (4.2 Å) is still too large for hydrogen bond formation. Probably Thr 141 is involved in the isomerization of the C3 carbon of EvsA. If mutations of T141Y and Y139T are made, it is possible that the situation could be

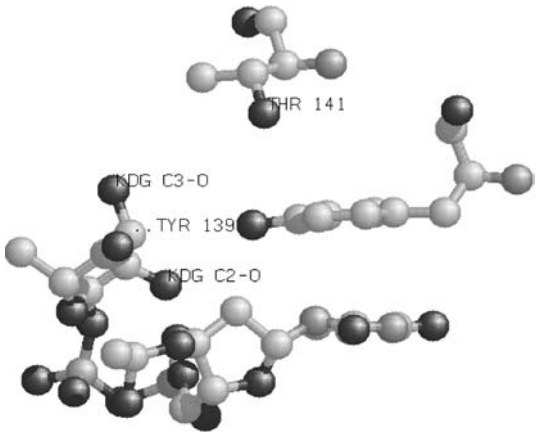

Fig. 12. Substrate and two residues from the active site of EvsA.

reversed. Tyr 141 could be hydrogen bonded to the hydroxyl group in the C3, thus preventing its isomerization, while Thr 139 would be at the correct distance to help the isomerization of the hydroxyl group in the C2. This means that the mutant might catalyze a 2,5-epimerization instead of a 3,5-epimerization.

CONCLUSIONS

We have discussed some of the key issues in predicting the structure and functions of proteins by homology recognition. The new tools described here are particularly useful in improving the identification of homologues and sequence-structure alignments. By exploiting information about protein 3D structure we can obtain a better understanding of the function of many proteins whose sequence are known.

ACKNOWLEDGMENTS

We thank Simon Lovell and Lucy Stebbings for reading the manuscript and Oliver W. Choroba for supplying useful information. RNM thanks the “Ministerio Español de Educación y Cultura” for financial support. KM is a Wellcome Trust Research Career Development Fellow.

References

1. Mizuguchi K, Deane CM, Blundell TL, Overington JP: **HOMSTRAD: a database of protein structure alignments for homologous families**. *Protein Sci* 1998, **7**:2469-2471.

2. Sali A, Blundell TL: **Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming**. *J Mol Biol* 1990, **212**:403-428.

3. a) Thompson JD, Plewniak F, Poch O: **BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs**. *Bioinformatics* 1999, **15**:87-88; b) Bahr A, Thompson JD, Thierry JC, Poch O: **BAliBASE (Benchmark Alignment data BASE): enhancements for repeats, transmembrane sequences and circular permutations**. *Nucleic Acids Res* 2001, **29**:323-326.

4. Bakker PIW, Bateman A, Burke DF, Miguel RN, Mizuguchi K, Shi J, Shirai H, Blundell TL: **HOMSTRAD: Adding sequence information to structure-based alignments of homologous protein families**. *Bioinformatics* 2001, **17**:in press.

5. Guda C, Scheeff ED, Bourne PE, Shindyalov IN: **A new algorithm for the alignment of multiple proteinstructures using Monte Carlo optimization**. *Pac Symp Biocomput* 2001, 275-286.

6. Overington JP, Donnelly D, Johnson MS, Sali A, Blundell TL: **Environment specific amino acid substitution tables: Tertiary templates and prediction of protein folds**. *Protein Science* 1991, **1**:216-226.
7. Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP: **JOY: protein sequence-structure representation and analysis**. *Bioinformatics* 1998, **14**:617-623.

8. Burke DF, Deane CM, Nagarajaram HA, Campillo N, Martin-Martinez M, Mendes J, Molina F, Perry J, Reddy BVB, Soares CM, Steward RE, Williams M, Carrondo MA, Blundell TL, Mizuguchi K: **An iterative structure-assisted approach to sequence alignment and comparative modelling**. *Proteins* 1999, **Suppl 3**:55-60.
9. Parker JS, Mizuguchi K, Gay NJ: **Family of proteins related to Spaetzle, the Toll receptor ligand, are encoded in the *Drosophila* genome**. *Proteins*, in press.

10. Shi J, Blundell TL, Mizuguchi K: **FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties**. *J Mol Biol* 2001, **310**:243-257.

11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

12. Lindahl E, Elofsson A: **Identification of related proteins on family, superfamily and fold level.** *J Mol Biol* 2000, **295**:613-625.

13. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J. Mol. Biol*. 1995, **247**:536-540.

14. Williams MG, Shirai H, Shi J, Nagendra HG, Mueller J, Mizuguchi K, Miguel RN, Lovell SC, Innis CA, Deane CM, Chen L, Campillo N, Burke DF, Blundell TL, Bakker PIW: **Sequence-Structure Homology Recognition by Iterative Alignment Refinement and Comparative Modelling**. *Proteins*, in press.

15. Thompson JD, Higgins DG, Gibson TJ: **Clustalw: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucl Acids Res* 1994, **22**:4673-4680.

16. Fischer D, Elofsson A, Rice D, Eisenberg D: **Assessing the performance of fold recognition methods by means of a comprehensive benchmark**. *Pac Symp Biocomput* 1996, 300-318.

17. Jones DT: **GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences**. *J Mol Biol* 1999, **287**:797-815.

18. Wageningen AMA, Kirkpatrick PN, Williams DH, Harris BR, Kershaw JK, Lennard NL, Jones M, Jones SJM, Solenberg PJ: **Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic**. *Chem Biol* 1998, **5**:155-162.

19. Williams DH, Bardsley B: **The Vancomycin Group of Antibiotics and the Fight against Resistant Bacteria**. *Angew Chem Int Ed* 1999, **38**:1172-1193.

20. Walker DR, Koonin EV: **SEALS: A System for Easy Analysis of Lots of Sequences.** *Intelligent Systems for Molecular Biology* 1997, **5**:333-339.

21. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices**. *J Mol Biol* 1999, **292**:195-202.

22. Rost B, Sander C: **Combining evolutionary information and neural networks to predict protein secondary structure**. *Proteins* 1994, **19**:55-72.

23. Mehta P, Heringa J, Argos P: **A simple and fast approach to prediction of protein secondary structure from multiple aligned sequences with accuracy above 70%**. *Protein Sci* 1995, **4**:2517-2525.

24. Frishman D, Argos P: **Knowledge-based secondary structure assignment**. *Proteins* 1995, **23**:566-579.

25. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **Jpred: A Consensus Secondary Structure Prediction Server**. *Bioinformatics* 1998, **14**:892-893.

26. Sayle RA, Milner-White EJ: **RasMol – Biomolecular graphics for all**. *Trends Biochem Sci* 1995, **20**:374-376.

27. Giraud M-F, Leonard GA, Field RA, Berlind C, Naismith JH: **RmlC, the third enzyme of dTDP-L- rhamnose pathway, is a new class of epimerase**. *Nature Struct Biol* 2000, **7**:398-402.

28. (a) Bairoch A: **The PROSITE dictionary of sites and patterns in proteins, its current status.** *Nucl Acids Res* 1993, **21**:3097-3103; (b) Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999.** *Nucl Acids Res* 1999, **27**:215-219.

29.(a) Henikoff S,Henikoff JG: **Automated assembly of proteins block for database searching.** *Nucl Acids Res* 1991, **19**:6565-6572; (b) Henikoff JG, Henikoff S, Pietrokovski S: **New features of the Vlocks Database servers**. *Nucl Acids Res* 1999, **27**:226-228.

30. Attwood TK, Flower DR, Lewis AP, Mabey JE, Morgan SR, Scordis P, Selley JN, Wright W: **PRINTS prepares for the new millennium**. *Nucl Acids Res* 1999, **27**: 220-225.

31. Bachinsky AG, Frolov AS, Naumochkin AN, Nizolenko LPh, Yarigin AA: **PROF_PAT 1.3: Updated database of patterns used to detect local similarities.** *Bioinformatics* 2000, **16**:358-366.

32. Srinivasan BN, Blundell TL: **An evaluation of the performance of an automated procedure for comparative modellingof protein tertiary structure**. *Protein Eng* 1993, **6**:501-512.

33. Sali A, Blundell TL: **Comparative Protein Modelling by Satisfaction of Spatial Restraints**. *J Mol Biol* 1993, **234**:779-815.

34. Deane CM, Kaas Q, Blundell TL: **SCORE: Predicting the core of protein models**. *Bioinformatics* 2001, **17**:541-550.

35. (a) Donate LE, Rufino SD, Canard LHJ, Blundell TL: **Conformational analysis and clustering of short and medium size loops connecting regular secondary structures. A database for modelling and prediction**. *Protein Sci* 1996, **5**:2600-2616; (b) Burke DF, Dean CM, Blundell TL: **A browsable and searchable web interface to the database of structurally based classification of loops – Sloop.** *Bioinformatics* 2000, **16**:513-519.

36. Deane CM, Blundell TL: **CODA: A combined algorithm for predicting the structurally variable regions of protein models**. *Protein Sci* 2001, **10**:599-612.

37. SYBYL is distributed by Tripos inc., 1699 South Hanley Road, St. Louis, MO 63144, USA.

38. Bower MJ, Cohen FE, Dunbrack RLJr: **Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool**. *J Mol Biol* 1997, **267**:1268-1282.

39. CELIAN: unpublished results.

40. Sutcliffe MJ, Haneef I, Carney D, Blundell TL: **Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures**. *Protein Eng*. 1987, **1**:377-384

41. a) Kikpatrick PN, Scaife W, Hallis TM, Liu H, Spencer JB, Williams DH: **Characterization of a sugar epimerase enzyme involved in the biosynthesis of a vancomycin-group antibiotic**. *Chem Commun* 2000, 1565-1566; b) Chen H, Thomas MG, Hubbard BK, Losey HC, Walsh CT, Burkart MD: **Deoxysugars in glycopeptide antibiotics: Enzymatic synthesis of TDP-L-epivancosamine in chloroeremomycin biosynthesis**. *PNAS* 2000, **97**:11942-11947.

42. InsightII is distributed by Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121, USA.

43. Laskowski RA, MacArthur MW, Moss DS, Thornton JM**: Procheck - A program to check the stereochemical quality of protein structures**. *J Appl Crystallogr* 1993, **26**:283-291.

44. Luthy R, Bowie JU, Eisenberg D: **Assessment of protein models with 3-dimensional profiles**. *Nature* 1992, **356**:83-85.

45. Sippl MJ: **Recognition of errors in the three-dimensional structure of proteins**. *Proteins* 1993, **17**:355-362.

46. Kraulis PJ: **MOLSCRIPT - A Program to produce both detailed and schematic plots of protein.** *J Appl Crystallogr* 1991, **24**:946-950.

47. Merritt EA, Murph M: **RASTER3D version-2.0 - A program for photorealistic molecular graphics**. *Acta Crystallogr D* 1994, **50**:869-873.

48. Nicholls A, Sharp K, Honig B: **Protein Folding and Association: Insights From the Interfacial and Thermodynamic Properties of Hydrocarbons**. *Proteins*, 1991, **11**:281ff.

Figure Legends

Fig. 1. From divergent evolution to protein 3D structure and function.

Fig. 2. Probabilities that a particular amino acid residue will not be substituted by any other residue type during evolution. The data were calculated from selected structure-based alignments in the HOMSTRAD database. Disulphide-bonded cysteine (C) and non-disulphide-bonded cysteine (J) residues are distinguished.

Fig. 3. Schematic representation of the steps followed in comparative modeling.

Fig. 4 Specificity-sensitivity curves of recognition performance at the family level using the test set provided by Dr. Elofsson.[12] Data other than that of FUGUE were kindly provided by Dr. Elofsson.

Fig. 5. Chemical reaction catalyzed by RmlC from *Amycolatopsis orientalis*.

Fig. 6. Prediction results from the FUGUE server using the sequence of EvsA as a query.

Fig. 7. Muconate lactonizing enzyme-like HOMSTRAD family aligned with two PROSITE motives and showing one of them.

Fig. 8. Sequence alignment of target EvsA with parent RmlC and formatted by JOY.

Fig. 9. a) Initial substrate analog of RmlC From *Salmonella typhimurium* place in the EvsA model. b) Modified substrate for EvsA. Figures generated by GRASP.[48]

Fig.10. a) Cartoon representation of the model of EvsA as a monomer. b) Representation of the active site of EvsA. Protein side chains (dark bonds) and substrate (light bonds) are illustrated. Hydrogen bonds are represented as dashed lines. c) Cartoon representation of the model of EvsA as a dimer.

Fig. 11. Schematic representation of the interactions at the EvsA active site. Hydrogen bonds are represented as dashed lines.

Fig. 12. Substrate and two residues from the active site of EvsA.