**COMMUNICATION**

# Three-dimensional modeling of the I-*Tev*I homing endonuclease catalytic domain, a GIY–YIG superfamily member, using NMR restraints and Monte Carlo dynamics

**Janusz M.Bujnicki[1,2], Piotr Rotkiewicz[3], Andrzej Kolinski[3] and Leszek Rychlewski[1,4]**

[1]Bioinformatics Laboratory, International Institute of Molecular and Cell Biology, ul. ks. Trojdena 4, 02-109 Warsaw, [3]Laboratory of Theory of Biopolymers, Department of Chemistry, Warsaw University, ul. Pasteura 1, 02-093 Warsaw and [4]BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznan, Poland

[2]To whom correspondence should be addressed. E-mail: iamb@bioinfo.pl

**Using a recent version of the SICHO algorithm for *in silico* protein folding, we made a blind prediction of the tertiary structure of the N-terminal, independently folded, catalytic domain (CD) of the I-*Tev*I homing endonuclease, a representative of the GIY–YIG superfamily of homing endonucleases. The secondary structure of the I-*Tev*I CD has been determined using NMR spectroscopy, but computational sequence analysis failed to detect any protein of known tertiary structure related to the GIY–YIG nucleases (Kowalski *et al.*, *Nucleic Acids Res.*, 1999, 27, 2115–2125). To provide further insight into the structure–function relationships of all GIY–YIG superfamily members, including the recently described subfamily of type II restriction enzymes (Bujnicki *et al.*, *Trends Biochem. Sci.*, 2000, 26, 9–11), we incorporated the experimentally determined and predicted secondary and tertiary restraints in a reduced (side chain only) protein model, which was minimized by Monte Carlo dynamics and simulated annealing. The subsequently elaborated full atomic model of the I-*Tev*I CD allows the available experimental data to be put into a structural context and suggests that the GIY–YIG domain may dimerize in order to bring together the conserved residues of the active site.**

*Keywords*: *ab initio* modeling/SICHO/structure-based function prediction/structure prediction

## Introduction

Knowledge about the three-dimensional (3D) structure of a protein is a key to comprehension and manipulation of its function. It may give clues, not apparent from the sequence alone, about homologs or analogs that share a catalytic mechanism or recognize the same ligand. Owing to technical difficulties and the high costs associated with experimental protein structure elucidation, it is often useful to predict (to model) the structure of a protein of interest and use it to guide other experiments on the laboratory bench. The success and utility of protein structure prediction have increased dramatically because of the multitude of sequences generated by the many genome sequencing projects. A wide range of theoretical approaches have been applied to this problem, the most reliable at present being those derived from homology modeling methods, which use experimentally determined protein structure to predict the conformation of another, evolutionarily related protein. With the rapid growth of protein sequence databases, many protein families are now known which exhibit clear similarity to the sequence of a protein of known tertiary structure. For other protein families, remote relationships can be inferred using more sensitive methods for alignment of sequence profiles (Rychlewski *et al.*, 2000) or sequence-to-structure threading (Jones *et al.*, 1999; Murzin, 1999; David *et al.*, 2000). However, in many cases a suitable homolog is not available (or not detectable) in the database, which means that the model has to be built *ab initio* (Koehl and Levitt, 1999; Sippl, 1999).

A problem in the *ab initio* protein structure prediction methodology is to search a vast conformational space efficiently. The existence of an astronomically large number of local energetic minima reduces tremendously the effectiveness of any of the *in silico* folding algorithms available today. Various models have been proposed that simplify the folding problem by reducing the number of degrees of freedom in the system and using primitive interaction potentials derived from analysis of known protein structures (Friesner and Gunn, 1996; Honig, 1999). Their efficiency is restricted mainly by the accuracy with which a simplified model can represent the protein and the ability of the potential to distinguish the native-like conformation from the many possible alternative structures. Another limitation of the methodology is that only low to moderate resolution structures can be generated, since the description of the protein chain is usually very coarse and specific interactions such as hydrogen bonds are not modeled by the simple potential used. Nevertheless, algorithms for reasonable reconstruction of full atomic detail from such sparse information, such as coordinates of C-α atoms or side-chain centers, have been developed (Feig *et al.*, 2000).

One approach to predicting the tertiary structure of a protein is to use cubic lattices to act as the restricted spaces in which the polypeptide chain can fold. Skolnick and co-workers have carried out a number of studies of folding of small and medium size proteins (~100 residues) using both lattice and off-lattice models via dynamic Monte Carlo methods and simulated annealing (Kolinski *et al.*, 1999). Their recently developed SICHO method employs a high-coordination lattice representation of the protein chain that incorporates a variety of potentials designed to produce protein-like behavior. It has been demonstrated that for representative proteins in each of the structural classes, it has been possible to achieve the correct tertiary fold using only secondary structure and a limited number of distance constraints. The secondary structure of a protein can be predicted from its sequence by using a variety of statistical methods (http://maple.bioc.columbia.edu/eva) or determined experimentally, for instance using NMR spectroscopy. The long-range contacts of individual residues or secondary structure elements can be inferred theoretically or determined experimentally and translated into geometrical constraints to define a constraint satisfaction problem used to resolve the 3D structure of a protein (Taylor, 1993). The power of such an approach lies in the possibility of observing interplay between experimentally derived restraints and theoretically predicted structure and to generate a consensus model.

Here, we describe a blind prediction of the tertiary structure of the N-terminal, independently folded, catalytic domain (CD) of the I-*Tev*I homing endonuclease (ENase), a representative of the GIY–YIG superfamily of deoxyribonucleases (Kowalski *et al.*, 1999). Homing ENases are enzymes encoded in introns or inteins. They recognize an extended sequence within an intronless gene and cut it, inducing a double strand break repair that leads to insertion of the intron (Belfort and Roberts, 1997; Jurica and Stoddard, 1999). Based on sequence comparisons they have been classified into four families characterized by the LAGLIDADG, GIY–YIG and H–N–H and His–Cys box motifs (Belfort and Perlman, 1995). Through structural comparisons it has been found that the H–N–H and His–Cys box enzymes, and also the non-specific nuclease from *Serratia* and phage T7 ENase VII, can be classified as a single superfamily, termed 'ββα-Me' to reflect the common secondary structure elements and the metal ion at the active site (Kuhlmann *et al.*, 1999).

The GIY–YIG superfamily is the only class of homing ENases for which high-resolution structures are not yet available. I-*Tev*I, the best studied GIY–YIG ENase, possesses a bipartite structure with separable catalytic (N-terminal) and DNA binding (C-terminal) domains separated by a flexible linker, similar to type IIS restriction enzymes, such as *Fok*I (Derbyshire *et al.*, 1997). Recently, the secondary structure of the I-*Tev*I CD has been determined using NMR spectroscopy (Kowalski *et al.*, 1999). It has been also shown that the GIY–YIG family includes the 3′ incision domain of the UvrC proteins (Kowalski *et al.*, 1999) and a subfamily of GGCGCC-specific type II restriction ENases (Bujnicki *et al.*, 2001). Nevertheless, the computational sequence analysis failed to detect any protein of known tertiary structure related to the GIY–YIG nucleases, suggesting that they may represent a novel fold or a significant modification of a known fold. In the absence of high-resolution structural model it is difficult to interpret the effect of mutation of putative catalytic residues and make generalizations about evolution of structure and function in widely diverged members of the superfamily. To provide further insight into structure–function relationship of all GIY–YIG nucleases, we incorporated the secondary and tertiary restraints from the NMR experiment (Kowalski *et al.*, 1999) and multiple sequence alignment in a reduced protein model minimized by Monte Carlo dynamics and simulated annealing.

## Methods

### *Database searches and sequence alignment*

The nr database and also genomic databases at NCBI were extensively screened using the PSI-BLAST algorithm (Altschul *et al.*, 1997), with the I-*Tev*I homing endonuclease sequence used as the basis for comparison. The full-length protein sequence alignment was constructed using the 'align sequences to profile' option of CLUSTALX (Thompson *et al.*, 1997) and the PSI-BLAST output as the starting point. The positions of gaps were adjusted to maintain continuity of secondary structure elements determined by NMR in I-*Tev*I (Kowalski *et al.*, 1999).

We also made an attempt to predict the tertiary structure of the I-*Tev*I ENase, and also other GIY–YIG nucleases, using various sequence-to-structure threading algorithms (available via the Metaserver interface at http://bioinfo.pl/meta), hoping to identify structurally characterized proteins of similar fold. However, none of the threading algorithms reported significant hits to any structure from the Protein Data Bank. Moreover, even the best hits reported were structurally dissimilar (data not shown), so we resorted to *ab initio* structure prediction.

### *Model building*

The multiple sequence alignment of the I-*Tev*I CD with other GIY–YIG nucleases was used as input for the blind tertiary structure prediction using a recently developed version of the SICHO program (Kolinski *et al.*, 1999; Skolnick *et al.*, 2000) with detailed derivations and methodology provided therein. Briefly, the procedure employs a 646 vector-based lattice protein model with a lattice spacing of 1.45 Å (Kolinski and Skolnick, 1998) and incorporates potentials reflecting short- and long-distance statistical preferences for secondary and tertiary structure. In the case of I-*Tev*I, potentials were weighted as previously described for small α/β-proteins (Kolinski *et al.*, 1999). Nine tertiary contacts read directly from the crude NMR model were used in the form of relatively strong conformational restraints. The following restraints were used: I5–A21, Y6–S20, Q7–G19, I8–V18, K9–Y17, G4–I64, I5–E63, Y6–E62 and Q7–L61 (Kowalski *et al.*, 1999). As a result, the restrained parts of the modeled structure did not move too far from the starting position. Sampling of conformational space was performed by the very efficient Replica Exchange Monte Carlo method (Gront *et al.*, 2000).

Twenty long independent simulations (of 10 replicas used in each run) starting from a fully extended initial conformation were carried out. Low-energy structures were then subject to a short isothermal Monte Carlo refinement at a low temperature below the folding transition. The structures exhibiting the lowest average energy during the isothermal calculations were assumed to represent the correct fold, according to the 'thermodynamic hypothesis', which our model tries to follow. The hypothesis says that the native conformations of proteins correspond to global minima of their free energy (Anfinsen, 1973). To construct a detailed model, the main chain representation was built from the side chain-only model based on local similarity to experimentally solved structures (Feig *et al.*, 2000). The all-atom refinement was carried out using GROMOS (Scott *et al.*, 1999) to improve local geometry and side chain packing.

## Results and discussion

### *Overall structure of the I-TevI CD*

The three-dimensional model of the I-*Tev*I CD (aa residues 1–94) was built as described in Methods, based on secondary and tertiary constraints derived from NMR analysis (Kowalski *et al.*, 1999) and multiple sequence alignment (Figure 1). Interestingly, the structures of lowest energy coming from different simulations exhibited the same common fold within range of resolution of the simplified model. The predicted structure consists of a single α/β domain with a three-stranded antiparallel β-sheet sandwiched between two α-helices, numbered α1 and α3 (Figure 2). Helix α2 assumes an unusual orientation, nearly perpendicular with respect to all other secondary structure elements. Helices α2 and α3 were unstable in the observed folding trajectory. This may suggest some conformational mobility in the native state. Alternatively, given the high sequence variability of the region of helix α2, such a result may reflect the adoption of a conformation which is evolutionarily variable, but well defined in individual, distinct structures (see also below).

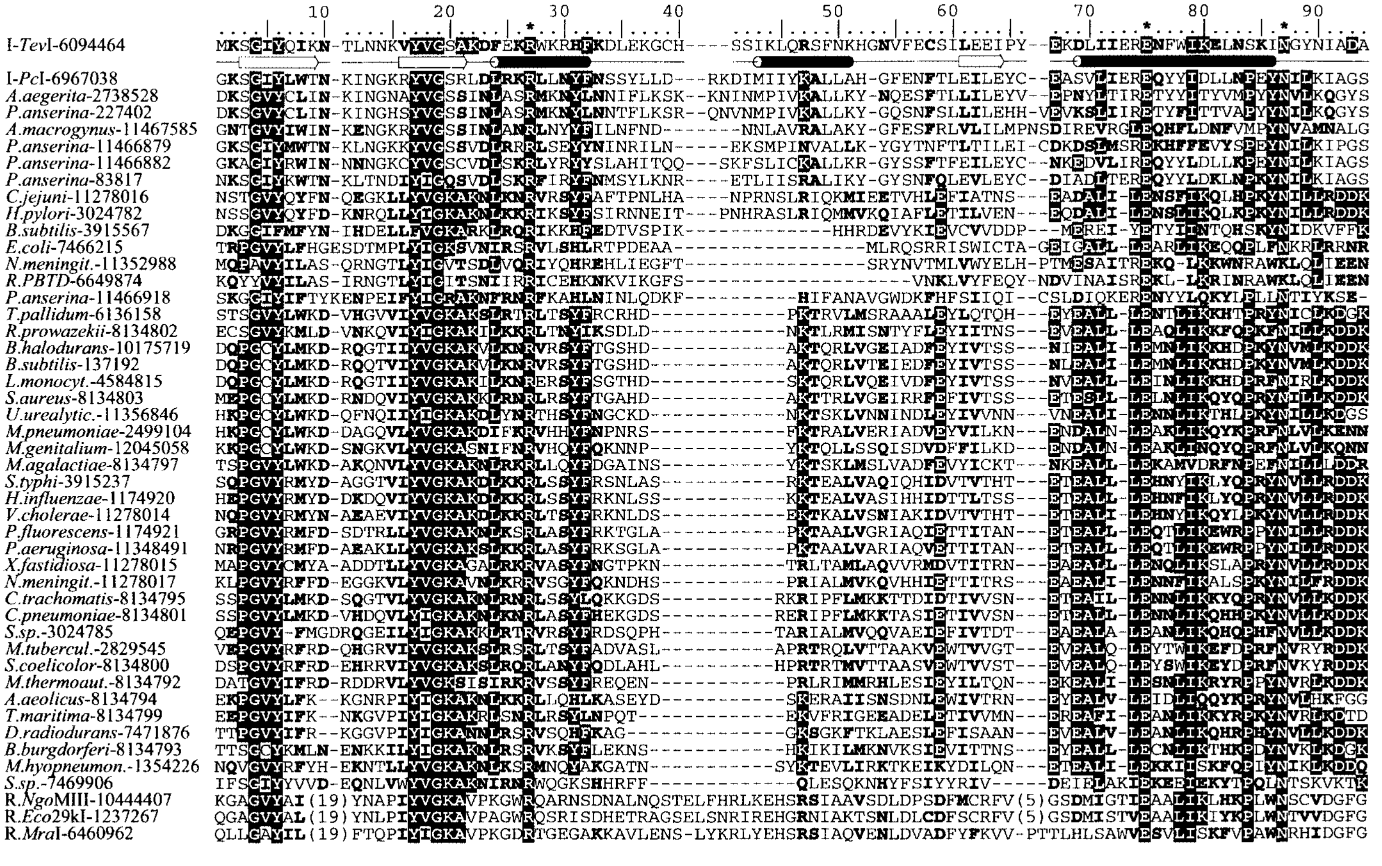

**Fig. 1**. Multiple sequence alignment of the GIY–YIG superfamily members with their Gene Identification numbers shown on the left. Residues conserved in ⩾50% of sequences are shown in black and residues with a similar physicochemical character in ⩾50% of sequences in gray. Numbers in parentheses indicate the length of sequence fragments omitted for the clarity of presentation. Secondary structure elements experimentally determined for the I-*Tev*I CD by Kowalski *et al.* (Kowalski *et al.*, 1999) are shown as cylinders (helices) and arrows (strands). Three putative catalytic residues are indicated by asterisks.

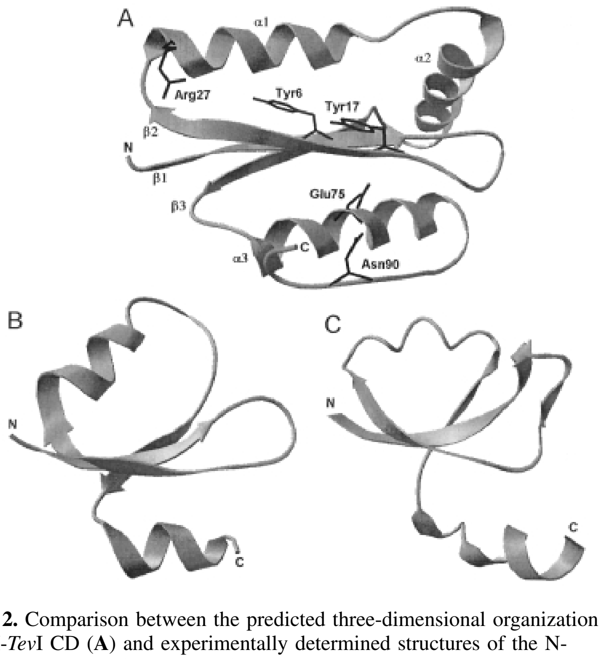

**Fig. 2.** Comparison between the predicted three-dimensional organization of the I-*Tev*I CD (**A**) and experimentally determined structures of the N-terminal domain of RNase HI from *Saccharomyces cerevisiae* (1qhk) (**B**) and the N-terminal domain of ribosomal protein L9 (1div) (**C**). Functionally important residues and secondary structure elements in the I-*Tev*I CD are labeled.

The topology of the β-sheet is identical with that reported by Kowalski *et al.* (Kowalski *et al.*, 1999), which indicates that no experiment-based tertiary constraints were violated by the folding algorithm. An analysis of the predicted structure of the I-*Tev*I CD using WHATCHECK (Hooft *et al.*, 1996) and VERIFY3D (Eisenberg *et al.*, 1997) indicates that the quality of the present model is acceptable. Bond angles and lengths were found to deviate normally from the mean standard bond angles (WHATCHECK *Z*-scores 1.466 and 0.941, respectively). No steric clashes were detected. Most importantly, according to the VERIFY3D algorithm, all residues along the entire polypeptide chain are compatible with the environment in which they were modeled. Even though the average value (0.22) is lower than for typical well-refined X-ray structures, it indicates that all structural elements, including solvent-exposed loops, assume a native-like arrangement, which suggests that the predicted topology is correct. Moreover, the initial NMR restraints were preserved in the final model, which indicates that the predicted topology is reliable.

The Protein Data Bank (PDB) was searched using the VAST server (Orengo *et al.*, 1997) for proteins or domains of proteins exhibiting similarity to the predicted structure of the I-*Tev*I CD. The results revealed two proteins with nearly identical fold, namely the N-terminal domain of ribosomal protein L9 (PDB entry 1div) and the N-terminal domain of RNase HI from *Saccharomyces cerevisiae (*1qhk). These proteins exhibit the β-β-α-β-α topology, that is they lack the counterpart of α2 in the I-*Tev*I structure. According to the multiple sequence alignment (Figure 1), the region of helix α2 is most variable in the GIY–YIG domain. It is tempting to speculate that those GIY–YIG nucleases, in which part of this region is deleted, exhibit an architecture identical with that of the two structurally characterized RNA-binding proteins. Interestingly, the type II GIY–YIG restriction ENases *Eco*29kI, *Ngo*MIII and *Mra*I possess a unique 16 aa insertion between strands β1 and β2 (Bujnicki *et al.*, 2001), which according to the model presented here, would be localized in the vicinity of the variable helix α2.

The structural similarity could have occurred by chance or it could reflect an extremely remote evolutionary relationship between the domains. No statistically significant sequence similarity between the I-*Tev*I CD and the two RNA-binding domains could be detected using algorithms either for iterative sequence database searches or sequence-structure threading (data not shown). However, examples are known of proteins that have similar structures and no detectable sequence similarity and sometimes even different active sites, despite functional similarities. For instance, we have recently analyzed significant structural similarity between the PD-(D/E)XK superfamily of deoxyribonucleases and the EndA family of tRNA splicing endonucleases, which despite the common fold use different surfaces to bind their nucleic acid substrates and possess dissimilar active sites, carrying out chemically distinct reactions (Bujnicki and Rychlewski, 2001).

### *Predicted active site and the cleavage mode of the I-TevI CD*

It has been suggested that the hallmark GIY and YVG sequence elements play a vital role in maintaining the integrity of the β-sheet, regardless of the potential role of the conserved Tyr residues in phosphodiester bond cleavage (Kowalski *et al.*, 1999). Our model agrees perfectly with this prediction, with only the Y17 side chain partially exposed to the solvent and positioned in the vicinity of other conserved residues, including R27, E75 and N90. Whereas I-*Tev*I mutants Y6A, G19A, R27A and E75A have no detectable catalytic activity, N90A and Y17A display a greatly reduced level of cleavage compared with the wild-type enzyme (Kowalski *et al.*, 1999).

The mechanism of cleavage of a phosphodiester bond is characterized by a general base that activates the attacking nucleophile, a Lewis acid that stabilizes the pentacovalent intermediate and a general acid that protonates the leaving group (Pingoud and Jeltsch, 1997). In I-*Tev*I R27 was proposed to function as the Lewis acid and E75 as a general base and a metal-binding residue (Kowalski *et al.*, 1999). Surprisingly, our model of the I-*Tev*I CD suggested that the GIY–YIG enzyme differs from other known nucleases with respect to the composition and proposed organization of the putative active site. In the PD-(D/E)XK, LAGLIDADG and ββαMe superfamilies of nucleases (Jurica and Stoddard, 1999; Aravind *et al.*, 2000), the invariant and partially conserved residues cluster towards the same side of the enzyme. However, in our model, Y17, E75 and N90 cluster together, whereas the indispensable R27 is localized more than 12 Å away. Importantly, in none of the alternative or intermediate models could all the conserved residues be clustered together without severe violation of the secondary structure constraints derived from the NMR experiment (data not shown). This suggests that R27 does not participate directly in the formation of the active site or, alternatively, that the GIY–YIG nuclease active site is formed *in trans* and includes R27 and E75 side chains from distinct polypeptides.

Previously, it has been argued that I-*Tev*I is a monomeric enzyme that binds its homing site and effects distant double-strand cleavage via a flexible hinge at a range of positions spanning two helical turns (Mueller *et al.*, 1995). The two-domain hinged monomer model of action was corroborated by the results of cleavage-site mapping and insertion mutagenesis (Bryk *et al.*, 1995), limited proteolysis experiments and mutagenesis of the proposed active site (Derbyshire *et al.*, 1997). However, the possibility that the GIY–YIG nucleases could function as dimers has never been ruled out. The type IIs restriction enzyme *Fok*I, a PD-(D/E)XK superfamily member, seemed to be the paradigm for a monomeric enzyme, which has only one catalytic center, but nevertheless makes a double-strand cut. *Fok*I, similarly to I-*Tev*I, is a bipartite enzyme, with two separate domains dedicated to DNA binding and catalysis. However, it was shown that the catalytic domains of *Fok*I must dimerize for DNA cleavage (Bitinaite *et al.*, 1998) and a model for the dimer of cleavage domain bound to the *Fok*I cleavage site was proposed (Wah *et al.*, 1998). In contrast, the LAGLIDADG nuclease PI-*Sce*I acts as a monomer, but its structure is characterized by a pseudo-two-fold symmetry and it possesses two similar active sites for separate cleavage of two strands of the target sequence (Christ *et al.*, 1999). Remarkably, in all LAGLIDADG nucleases, including those possessing duplicated catalytic domains and the *bona fide* homodimeric enzymes, each of the catalytic centers is composed of side chains from two separate domains, together making up an intertwined 'ying-yang' motif (Christ *et al.*, 1999).

Recently, it has been found that type II restriction enzymes *Eco*29kI, *Mra*I and *Ngo*MIII belong to the GIY–YIG superfamily (Bujnicki *et al.*, 2001). The finding of dimeric GIY–YIG nucleases supports our prediction that, in analogy with the PD-(D/E)XK superfamily members, i.e. dimeric type II ENases and transiently dimerizing type IIs ENases, two catalytic domains of I-*Tev*I may need to form a temporary complex with one target sequence to exert the double strand cleavage. Our prediction that the active site of I-*Tev*I is assembled *in trans* suggests that I-*Tev*I mutants R27A and E75A should complement each other. It would be interesting to test whether the dimer with only one functionally active site is capable of nicking only one of the strands, suggesting a fixed orientation of the two catalytic domains with respect to the TRD bound to the homing site or whether the I-*Tev*I CD is flexible enough to make a double strand cut.

### *Conclusions*

The structural model of the I-*Tev*I CD presented in this paper suggests that GIY–YIG nucleases are structurally similar to a domain identified in nucleic acid-binding proteins RNase HI and ribosomal protein L9. Based on the predicted structure, we propose a 'ying-yang' model of the GIY–YIG nuclease active site, which implies that dimerization of the catalytic domain is needed for the cleavage reaction to occur. It is worth emphasizing that the structure of the I-*Tev*I CD could not be predicted using 'standard' tools for computational sequence analysis, including threading programs. Therefore, our analysis demonstrates the value of algorithms for *ab initio* structure prediction in inferring the details of the molecular function of proteins, for which experimental data are insufficient to provide a satisfactory picture of structure–function relationships. It will be interesting to compare the presented model with the experimentally solved three-dimensional structure of I-*Tev*I and to test the hypothesis of the 'ying-yang' model of the GIY–YIG active site by site-directed mutagenesis of I-*Tev*I or related restriction enzymes.

## References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Anfinsen,C.B. (1973) *Science*, **181**, 223–230.

Aravind,L., Makarova,K.S. and Koonin,E.V. (2000) *Nucleic Acids Res.*, **28**, 3417–3432.

Belfort,M. and Perlman,P.S. (1995) *J. Biol. Chem.*, **270**, 30237–30240.

Belfort,M. and Roberts,R.J. (1997) *Nucleic Acids Res.* **25**, 3379–3388.

Bitinaite,J., Wah,D.A., Aggarwal,A.K. and Schildkraut,I. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 10570–10575.

Bryk,M., Belisle,M., Mueller,J.E. and Belfort,M. (1995) *J. Mol. Biol.*, **247**, 197–210.

Bujnicki,J.M. and Rychlewski,L. (2001) *Protein Sci.*, **10**, 656–660.

Bujnicki,J.M., Radlinska,M. and Rychlewski,L. (2001) *Trends Biochem. Sci.*, **26**, 9–11.

Christ,F., Schoettler,S., Wende,W., Steuer,S., Pingoud,A. and Pingoud,V. (1999) *EMBO J.*, **18**, 6908–6916.

David,R., Korenberg,M.J. and Hunter,I.W. (2000) *Pharmacogenomics*, **1**, 445–455.

Derbyshire,V., Kowalski,J.C., Dansereau,J.T., Hauer,C.R. and Belfort,M. (1997) *J. Mol. Biol.*, **265**, 494–506.

Eisenberg,D., Luthy,R. and Bowie,J.U. (1997) *Methods Enzymol.*, **277**, 396–404.

Feig,M., Rotkiewicz,P., Kolinski,A., Skolnick,J. and Brooks,C.L. (2000) *Proteins*, **41**, 86–97.

Friesner,R.A. and Gunn,J.R. (1996) *Annu. Rev. Biophys. Biomol. Struct.*, **25**, 315–342.

Gront,D., Kolinski,A. and Skolnick,J. (2000) *J. Chem. Phys.*, **113**, 5065–5071.

Honig,B. (1999) *J. Mol. Biol.*, **293**, 283–293.

Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) *Nature*, **381**, 272.

Jones,D.T., Tress,M., Bryson,K. and Hadley,C. (1999) *Proteins*, **37**, 104–111.

Jurica,M.S. and Stoddard,B.L. (1999) *Cell Mol. Life Sci.*, **55**, 1304–1326.

Koehl,P. and Levitt,M. (1999) *Nature Struct. Biol.*, **6**, 108–111.

Kolinski,A. and Skolnick,J. (1998) *Proteins*, **32**, 475–494.

Kolinski,A., Rotkiewicz,P., Ilkowski,B. and Skolnick,J. (1999) *Proteins*, **37**, 592–610.

Kowalski,J.C., Belfort,M., Stapleton,M.A., Holpert,M., Dansereau,J.T., Pietrokovski,S., Baxter,S.M. and Derbyshire,V. (1999) *Nucleic Acids Res.*, **27**, 2115–2125.

Kuhlmann,U.C., Moore,G.R., James,R., Kleanthous,C. and Hemmings,A.M. (1999) *FEBS Lett.*, **463**, 1–2.

Mueller,J.E., Smith,D., Bryk,M. and Belfort,M. (1995) *EMBO J.*, **14**, 5724–5735.

Murzin,A.G. (1999) *Proteins*, **Suppl.3**, 88–103.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.

Pingoud,A. and Jeltsch,A. (1997) *Eur. J. Biochem.*, **246**, 1–22.

Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) *Protein Sci.*, **9**, 232–241.

Scott,W.R.P., Hunenberger,P.H., Tironi,I.G., Mark,A.E., Billeter,S.R., Fennen,J., Torda,A.E., Huber,T., Kruger,P. and van Gunsteren,W.F. (1999) *J. Phys. Chem.*, **103**, 3596–3607.

Sippl,M. (1999) *Structure*, **7**, R81-R83.

Skolnick,J., Kolinski,A. and Ortiz,A. (2000) *Proteins*, **38**, 3–16.

Taylor,W.R. (1993) *Protein Eng.*, **6**, 593–604.

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4876–4882.

Wah,D.A., Bitinaite,J., Schildkraut,I. and Aggarwal,A.K. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 10564–10569.

*Received February 1, 2001; revised July 4, 2001; accepted July 18, 2001*