Protein Engineering vol.13 no.7 pp.453–457, 2000

COMMUNICATION

Constructing side chains on near-native main chains for *ab initio* protein structure prediction

Ram Samudrala[1,2], Enoch S.Huang[3], Patrice Koehl[1] and Michael Levitt[1]

[1]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305 and [3]Cereon Genomics, 45 Sidney Street, Cambridge, MA 02139, USA

[2]To whom correspondence should be addressed.
E-mail: ram@csb.stanford.edu

Is there value in constructing side chains while searching protein conformational space during an *ab initio* simulation? If so, what is the most computationally efficient method for constructing these side chains? To answer these questions, four published approaches were used to construct side chain conformations on a range of near-native main chains generated by *ab initio* protein structure prediction methods. The accuracy of these approaches was compared with a naive approach that selects the most frequently observed rotamer for a given amino acid to construct side chains. An all-atom conditional probability discriminatory function is useful at selecting conformations with overall low all-atom root mean square deviation (r.m.s.d.) and the discrimination improves on sets that are closer to the native conformation. In addition, the naive approach performs as well as more sophisticated methods in terms of the percentage of χ_1 angles built accurately and the all-atom r.m.s.d., between the native and near-native conformations. The results suggest that the naive method would be extremely useful for fast and efficient side chain construction on vast numbers of conformations for *ab initio* prediction of protein structure.

Keywords: conditional probability/discriminatory function/knowledge-based/protein structure prediction/side chain construction

Introduction

Methods to predict the three-dimensional conformation of a protein from its amino acid sequence are maturing to a point where conformations ≤4.0 Å C_α root mean square deviation (r.m.s.d.) can be generated for small proteins or fragments of a protein (Mumenthaler and Braun, 1995; Park and Levitt, 1996; Pedersen and Moult, 1997; Huang *et al.*, 1998b; Koehl and Levitt, 1999; Lee *et al.*, 1999; Moult *et al.*, 1999; Orengo *et al.*, 1999; Ortiz *et al.*, 1999; Osguthorpe, 1999; Samudrala *et al.*, 1999; Simons *et al.*, 1999). Some of these methods generally search protein conformational space using only the main chain or C_α trace, ignoring side chains entirely or approximating their conformation using one or two positions in coordinate space (Sun, 1993; Park and Levitt, 1996; Simons *et al.*, 1999). This is done to make the computations, i.e. generating a new conformation and calculating its energy, more tractable. The corresponding scoring functions used to distinguish native-like conformations from non-native ones are devised to handle this approximation in the representation of protein structure (Park and Levitt, 1996; Park *et al.*, 1997; Simons *et al.*,1999).

This approximation reduces atomic detail and leads to certain information being ignored, such as the side chain–side chain and side chain–main chain atom interactions. However, considering the interactions of side chain atoms with other atoms in the environment has been shown to help discriminatory functions in choosing near-native conformations in the sample space more accurately (Samudrala and Moult, 1998a). Given the intractability in searching protein conformational space with an all-atom representation, a two-step procedure in which the search initially focuses on the main chain and side chains are added before the conformations are evaluated would be very useful. The method for side chain construction must be computationally efficient and as accurate as possible, considering that the main chains generated will be fairly distant (≥2.0 Å C_α r.m.s.d.) from the native conformation.

In this work, we first constructed side chains using four previously published methods (Levitt, 1992; Koehl and Delarue, 1994; Bower *et al.*, 1997; Samudrala and Moult, 1998b) on four proteins with a varying number of near-native (≤4.0 Å C_α r.m.s.d.) conformations generated by two *ab initio* protein structure prediction methods (Park and Levitt, 1996; Simons *et al.*, 1997). We compared the performance of these approaches with a naive approach that simply uses the side chain rotamer most frequently observed in protein structures. We then used a residue-specific all-atom conditional probability discriminatory function (RAPDF) to select the lowest all-atom conformations generated by each of the methods and compared the discrimination results obtained using all-atom information with those obtained using only main chain information. The implications of these results for *ab initio* protein structure prediction are discussed.

Methods

Details of proteins and near-native main chains used for side chain construction

Table I gives the details of the four proteins that were selected for evaluating side chain construction. The proteins had been used for two *ab initio* protein structure prediction studies where only the main chain information was used to explore the conformational space (Park and Levitt, 1995; Simons *et al.*, 1997). These methods generate conformations with C_α r.m.s.d. between 1.5 and 12.0 Å for these proteins and a subset of conformations ≤4.0 Å C_α r.m.s.d. were chosen for this study.

The method of Simons *et al.* starts with an extended polypeptide chain and uses Monte Carlo moves with simulated annealing in torsion angle space to generate compact conformations that are favored by a knowledge-based Bayesian scoring function. The move set for each residue is based on a library of fragments from a database of unrelated protein structures with similar local sequences (Simons *et al.*, 1997).

The method of Park and Levitt builds all main chains using four discrete (ϕ, ψ) values. The native secondary structure is

Table I. Details of the four proteins and sets of near-native conformations used for this study

Protein[a]	PDB code (chain)	Number of residues	Resolution (Å)	*R* value	Experimental reference	Number of conformations	R.m.s.d. (Å)	Generation method
Protein A	1fc2 (C)	43	2.8	0.24	Deisenhofer, 1981	177	3.7	Simons *et al.*, 1997
Homeodomain	1hdd (C)	57	2.8	0.22	Kissinger *et al.*, 1990	32	3.5	Simons *et al.*, 1997
434 Repressor	1r69	63	2.0	0.19	Mondragon *et al.*, 1989	97	3.3	Park and Levitt, 1995
Ubiquitin	1ubq	76	1.8	0.18	Vijay-Kumar *et al.*, 1987	19	3.2	Park and Levitt, 1995

[a]These proteins were used for *ab initio* protein structure prediction studies where only the main chain information is used to generate conformations. A subset of conformations with C_α r.m.s.d. ≤4.0 Å were chosen and the mean C_α r.m.s.d. is shown.

fixed and designated residues are permitted to explore all possibilities of the four (ϕ, ψ) values in a combinatorial fashion (Park and Levitt, 1996). This method generates conformations with lower C_α r.m.s.d. than the method of Simons *et al.*, but is more sensitive to knowledge of the exact protein secondary structure. The original *ab initio*-generated coordinates for the two methods are available from the Decoys 'R' Us database at <http://dd.stanford.edu>.

Residue-specific all-atom conditional probability discriminatory function (RAPDF)

We use an all-atom distance-dependent conditional probability-based discriminatory function to calculate the probability of a native structure, given a set of distances between pairs of atoms. A full description can be found in Samudrala and Moult (1998a). Briefly, the required probabilities are compiled by counting frequencies of distances between pairs of atom types in a database of protein structures. All non-hydrogen atoms are considered and the description of the atoms is residue specific, i.e. the C_α of an alanine is different from the C_α of a glycine. This results in a total of 167 atom types. We divide the distances observed into 1.0 Å bins ranging from 3.0 to 20.0 Å. Contacts between atom types in the 0.0–3.0 Å range are placed in a separate bin, resulting in total of 18 distance bins.

We compile tables of scores s proportional to the negative log conditional probability that we are observing a native conformation given an interatomic distance d for all possible pairs of the 167 atom types, a and b, for the 18 distance ranges, $P(C|d_{ab})$:

$$s(d_{ab}) = -\ln \frac{P(d_{ab}|C)}{P(d_{ab})} \propto -\ln P(C|d_{ab}) \quad (1)$$

where $P(d_{ab}|C)$ is the probability of observing a distance d between atom types a and b in a correct structure and $P(d_{ab})$ is the probability of observing such a distance in any structure, correct or incorrect. The required ratios $P(d_{ab}|C)/P(d_{ab})$ are obtained as follows:

$$\frac{P(d_{ab}|C)}{P(d_{ab})} = \frac{N(d_{ab})/\sum_d N(d_{ab})}{\sum_{ab} N(d_{ab})/\sum_d \sum_{ab} N(d_{ab})} \quad (2)$$

where $N(d_{ab})$ is the number of observations of atom types a and b in a particular distance bin d, $\Sigma_d N(d_{ab})$ is the number of a–b contacts observed for all distance bins, $\Sigma_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atoms types a and b in a particular distance bin d and $\Sigma_d \Sigma_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atom types a and b summed over all the distance bins d. No intra-residue distances are included in the summation. The tables of scores are compiled from a set of 312 unique folds from the SCOP database (Hubbard *et al.*, 1997).

Table II. Rotamer side chain χ angles (°) used in the naive approach

Amino acid	χ_1	χ_2	χ_3	χ_4
Cysteine	–58			
Serine	59			
Threonine	–59			
Valine	–177			
Asparagine	–60	0		
Aspartic acid	–57	0		
Histidine	–60	90		
Isoleucine	–57	163		
Leucine	–57	180		
Phenylalanine	–56	90		
Tryptophan	–58	90		
Tyrosine	–60	90		
Glutamic acid	–56	180	0	
Glutamine	–60	180	0	
Methionine	–58	–60	–60	
Arginine	–58	180	180	180
Lysine	–60	180	180	180

A naive approach for side chain construction

The naive approach simply constructs side chains based on the most frequently observed rotamer value in a database of protein structures (Table II). The particular values used were generated by the program mutate by R.Read (personal communication).

Comparison with previously published approaches: scgen, scmf, scwrl and segmod

Four previously published approaches were selected for comparison with the naive method. The approaches were primarily chosen because of their computational speed, their widespread use and their diversity in terms of methodology applied: (i) scgen uses the all-atom scoring function described above to select the lowest scoring rotamer from a discrete library considering interactions between the side chain atoms and the local main chain (Samudrala and Moult, 1998b); (ii) scmf uses self-consistent mean-field theory to position rotamers in conjunction with a van der Waals potential (Koehl and Delarue, 1994); (iii) scwrl uses a main chain dependent rotamer library to position side chains and minimizes the steric clashes (Bower *et al.*, 1997); and (iv) segmod pastes in side chain conformers directly from a structural database, using a Boltzmann-weighted probability to choose the conformation in the context of the main chain and the side chains already positioned (Levitt, 1992). All methods assume a fixed main chain at the time of side chain placement.

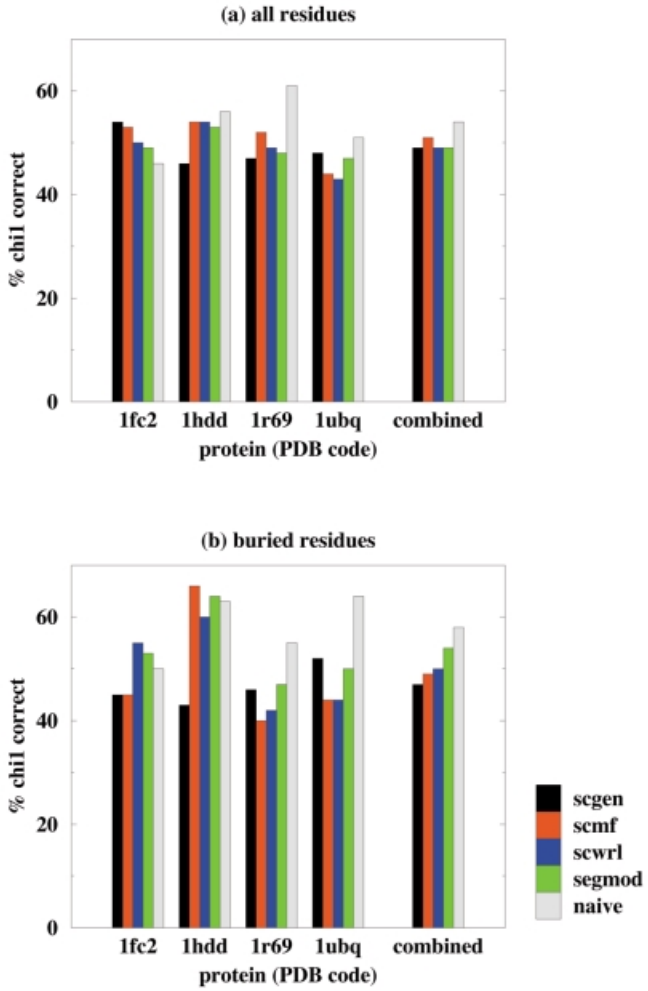

Fig. 1. Accuracy of side chain construction by five different methods for the four different sets of main chains. For each set, the horizontal bars show the percentage of χ_1 angles built correctly (within 40° of the native value) for all residues (**a**) and for buried residues (**b**). The 'combined' column gives the percentage over all the residues for all four sets. All methods, including the naive approach, build side chains with surprisingly similar accuracy.

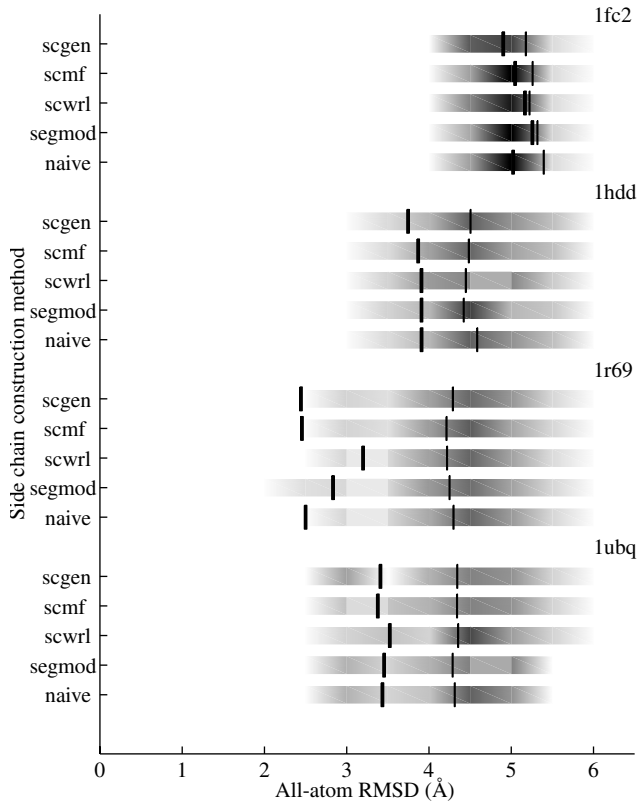

Fig. 2. Distribution of all-atom r.m.s.d.s for the different methods for the four sets of near-native conformations and evaluation of discrimination using an all-atom scoring function. The r.m.s.d. distributions are shown by shaded horizontal bars, with the intensity of shading indicating the fraction of conformations observed at a particular r.m.s.d.. The discrimination by the all-atom function is indicated by a thick vertical line and the discrimination using only main chain information is indicated by a thin vertical line. The all-atom r.m.s.d. distributions for a given set of near-natives are similar regardless of the method used and the naive method appears to perform as well as any of the more sophisticated methods. The discrimination using all-atom information is also similar between the sets generated by the different methods, with the function selecting conformations closest to the native conformations on sets with better all-atom r.m.s.d. distributions (1hdd, 1r69 and 1ubq). The selection is consistently better when all-atom information is taken into account.

Evaluating side chain placement accuracy

To evaluate side chain placement accuracy, we determine the percentage of χ angles constructed within $\pm 40°$ of the values observed in the native conformation. We also use the all-atom r.m.s.d. between the near-native conformations and the experimental conformation, which is calculated using the equation

$$\sqrt{\frac{\sum_{i=1}^{N}(\Delta x_i^2+\Delta y_i^2+\Delta z_i^2)}{N}} \qquad (3)$$

where Δx_i, Δy_i and Δz_i are distances in Cartesian space between N corresponding atoms. Coordinate superposition is performed using the program align (Satow *et al.*, 1986; McLachlan, 1979).

Designation of buried residues

Computation of solvent accessibility for each side chain was performed using the software naccess (by S.J.Hubbard and J.M.Thornton of University College, London). Side chains with relative solvent exposure of $\leqslant 20\%$ were considered to be buried.

Minimization of conformations

Since different methods have different protocols for generating side chains and since packing influences can have an effect on side chain accuracy, all conformations were minimized for 200 steps using encad (Levitt and Lifson, 1969; Levitt, 1974, 1983; Levitt *et al.*, 1995) as a means of 'normalizing' the side chain models to remove severe steric strain. The all-atom r.m.s.d. between the minimized and unminimized conformations on average for each set is <0.1 Å.

Results and discussion

Naive method does as well as more sophisticated methods in terms of percentage χ_1 accuracy

Figure 1 shows the accuracy of side chain construction by the five different methods on our test sets. All methods, including the naive approach, build side chains with similar accuracy (50–60%) in terms of the percentage of χ_1 angles within 40° for all residues and for buried residues in the core. The trend is identical even when both χ_1 and χ_2 angles are considered (with accuracies ranging from 40 to 50%).

The side chain construction accuracy for the naive approach is constant, no matter how distant the main chain is from the native structure, i.e. even on the native main chain the naive approach remains 50–60% accurate. The results for the accuracy of side chain construction for scmf, scwrl and segmod on the native main chains of the four proteins have been published (Huang *et al.*, 1998a) and are ~70% for the χ_1 angles for all residues and ~80% for the χ_1 angles in buried residues. The values for scgen are similar to these values.

Even though the results are similar, three of the methods (scwrl, segmod and the naive approach) rely predominantly on the knowledge base on protein structures whereas the other two (scgen and scmf) perform selection of side chains based on energetic criteria.

In previous work it has been shown that the percentage χ_1 accuracy has a theoretical upper limit of about 60% given the steric constraints imposed by the near-native main chains for this set of conformations (Huang *et al.*, 1998a). This combined with the results from Figure 1 would indicate that the more sophisticated methods mimic the naive approach on main chains where the C_α r.m.s.d. ranges from 1.5 to 4.0 Å.

Distribution of all-atom r.m.s.d.s and discrimination by an all-atom scoring function is similar for all methods

Although the results from Figure 1 are interesting, they expose a limitation in determining the accuracy of side chain construction using the percentage of χ_1 angles as a gauge on near-native main chains. This is because this measure does not account for the variance in the main chain conformations. Irrespective of how well the main chain is modeled, the χ_1 angles will always remain the same: the percentage χ_1 accuracy of the naive method on the native main chain and on a main chain that is, say, 10.0 Å away from the native conformation will be identical. A measure such as all-atom r.m.s.d. takes into account the variance in both the main chains and the side chains and the expectation would be that the sophisticated methods would produce better all-atom r.m.s.d.s since they are designed to perform better on main chains that closer to the native conformation. We therefore compute the all-atom r.m.s.d. for each set of conformations generated by the different methods to determine whether this expectation is true.

Figure 2 illustrates the distribution of the all-atom r.m.s.d.s for the five methods using a 'gel graph', where the limits of the horizontal bar indicate the all-atom r.m.s.d. range and the density of shading indicates the fraction of conformations observed at a particular r.m.s.d. Again, the naive method generates conformations with r.m.s.d. distributions that are as good as those observed for the more sophisticated methods.

This figure also shows the selection of the best scoring conformation using an all-atom conditional probability based scoring function (thick vertical bar). The all-atom function generally selects proteins that are closer to the native conformations (lower end of the r.m.s.d. range) for the sets with the good all-atom r.m.s.d. ranges (1hdd, 1r69 and 1ubq) regardless of the method used for side chain construction.

Ignoring side chain information results in worse discrimination

It has been shown before that taking side chain interactions into account leads to better discrimination (Samudrala and Moult, 1998a). For this particular set, using the all-atom scoring function to discriminate using only main chain information leads to an average selection accuracy that is worse by about 0.8 Å all-atom r.m.s.d. over all the sets (Figure 2). The discrimination is consistently better when all-atom information is taken into account.

Implications for protein structure prediction

Our work does not attempt an exhaustive comparison of side chain methods. Rather, the goal was to compare a naive approach, based on using the most frequently observed rotamer in known protein structures, with a set of more sophisticated side chain construction methods and to determine the utility of side chain construction on near-native main chains. We have found that the simple naive approach performs as well as the more sophisticated methods.

The all-atom function does better at discriminating near-native conformations on main chains that are closer to the native conformation (Figure 2) and side chain information is indeed important to achieve this discrimination even at low resolution. Thus current *ab initio* methods that use reduced representations of protein would be better off building side chains in the best manner possible. Given that millions or billions of conformations are generated in an *ab initio* simulation (Samudrala *et al.*, 1999; Xia *et al.*, 2000), taking side chains into account using the inexpensive naive approach appears to be a good trade-off between computation time and accuracy.

It is likely that as the conformations become closer to the native ($\leqslant$2.0 Å C_α r.m.s.d.), side chain construction by the more sophisticated methods will have a greater impact on discrimination. Although current *ab initio* methods do not sample conformations to this resolution, this suggests a future approach for side chain construction in *ab initio* prediction: (i) Construct side chains on all or a large subset of the main chains using the naive approach to produce detailed all-atom models; (ii) filter using an all-atom scoring function to give a low scoring subset; and (iii) for this subset, build side chains using one of the more sophisticated methods described previously.

The results obtained here used 'vanilla' versions of the side chain modeling software. Improvements to the more sophisticated methods have been published (Dunbrack, 1999). Similarly, the naive approach can be refined as the knowledge base of known protein structures becomes larger, which may lead to more accurate all-atom models and consequently more accurate discrimination.

Acknowledgements

This work was supported in part by a Burroughs Wellcome Fund Fellowship from the Program in Mathematics and Molecular Biology to Ram Samudrala, a Jane Coffin Childs Memorial Fund Fellowship to Enoch Huang and NIH Grant GM 41455 to Michael Levitt. Patrice Koehl acknowledges support from the Union Internationale Contre le Cancre (UICC). We thank Kim Simons and David Baker for providing us with their decoy set.

References

Bower,M., Cohen,F. and Dunbrack,R. (1997) *J. Mol. Biol*., **267**, 1268–1282.
Diesenhofer,J. (1981) *Biochemistry*, **20**, 2361–2370.
Dunbrack,R. (1999) *Proteins: Struct. Funct. Genet*., **S3**, 81–87.
Huang,E., Koehl,P., Levitt,M., Pappu,R. and Ponder,J. (1998a) *J. Mol. Biol*., **33**, 204–217.
Huang,E., Samudrala,R. and Ponder,J. (1998b) *Protein Sci*., **7**, 1998–2003.
Hubbard,T., Murzin,A., Brenner,S. and Chothia,C. (1997) *Nucleic Acids Res*., **25**, 236–239.
Kissinger,C.R., Liu,B.S., Martin-Blanco,E. Kornberg,T.B. and Pablo,C.O. (1990) *Cell*, **63**, 579–590.
Koehl,P. and Delarue,M. (1994) *J. Mol. Biol*., **239**, 249–275.
Koehl,P. and Levitt,M. (1999) *Nature Struct. Biol*., **6**, 108–111.

Lee,J., Liwo,A., Ripoll,D., Pillardy,J. and Scheraga,J. (1999) *Proteins: Struct. Funct. Genet*., **S3**, 204–208.
Levitt,M. (1974) *J. Mol. Biol*., **82**, 393–420.
Levitt,M. (1983) *J. Mol. Biol*., **168**, 595–620.
Levitt,M. (1992) *J. Mol. Biol*., **226**, 507–533.
Levitt,M. and Lifson,S. (1969) *J. Mol. Biol*., **46**, 269–279.
Levitt,M., Hirshberg,M., Sharon,R. and Daggett,V. (1995) *Comput. Phys. Commun*., **91**, 215–231.
McLachlan,A. (1979) *J. Mol. Biol*., **128**, 49–79.
Mondragon,A., Subbiah,S., Almo,S.C., Drottar,M. and Harrison,S.C. (1989) *J. Mol. Biol*., **205**, 189–200.
Moult,J., Hubbard,T., Fidelis,K. and Pedersen,J. (1999) *Proteins: Struct. Funct. Genet*., **S3**, 2–6.
Mumenthaler,C. and Braun,W. (1995) *Protein Eng*., **4**, 863–871.
Orengo,C., Bray,J., Hubbard,T., LoConte,L. and Sillitoe,J. (1999) *Proteins: Struct. Funct. Genet*., **S3**, 149–170.
Ortiz,A., Kolinkski,A., Rotkiewicz,P., Ilkowski,B. and Skolnick,J. (1999) *Proteins: Struct. Funct. Genet*., **S3**, 177–185.
Osguthorpe,D. (1999) *Proteins: Struct. Funct. Genet*., **S3**, 186–193.
Park,B., Huang,E. and Levitt,M. (1997) *J. Mol. Biol*., **266**, 831–846.
Park,B. and Levitt,M. (1995) *J. Mol. Biol*., **249**, 493–507.
Park,B. and Levitt,M. (1996) *J. Mol. Biol*., **258**, 367–392.
Pedersen,J.T. and Moult,J. (1997) *J. Mol. Biol*., **269**, 240–259.
Samudrala,R. and Moult,J. (1998a) *J. Mol. Biol*., **275**, 895–916.
Samudrala,R. and Moult,J. (1998b) *Protein Eng*., **11**, 991–997.
Samudrala,R., Xia,Y., Huang,E. and Levitt,M. (1999) *Proteins: Struct. Funct. Genet*., **S3**, 194–198.
Satow,Y., Cohen,G., Padlan,E. and Davies,D. (1986) *J. Mol. Biol*., **190**, 593–604.
Simons,K., Kooperberg,C., Huang,E. and Baker,D. (1997) *J. Mol. Biol*., **268**, 209–225.
Simons,K., Bonneau,R., Ruczinski,I. and Baker,D. (1999) *Proteins: Struct. Funct. Genet*., **S3**, 171–176.
Sun,S. (1993) *Protein Sci.,* **2**, 762–785.
Vijay–Kumar,S., Bugg,C.E. and Cook,W.J. (1987) *J. Mol. Biol.*, **194**, 531–544.
Xia,Y., Huang,E.S., Levitt,M. and Samudrala,R. (2000) *J. Mol. Biol*., **300**, 171–185.

Received January 4, 2000; revised April 1, 2000; accepted May 2, 2000