A structural explanation for the twilight zone of protein sequence homology

Su Yun Chung[1] and S Subbiah[2†]

Homology modeling of protein structures as a function of sequence breaks down at the twilight zone limit of sequence identity between the template and target proteins. Our results suggest that protein sequences that have diverged from a common ancestor beyond the twilight zone may adopt side-chain interactions that are very different from those endowed by the ancestral sequence.

Addresses: [1]Department of Biochemistry Uniformed Services, University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA and [2]Department of Structural Biology, Stanford University, School of Medicine, Fairchild Building, Stanford, CA 94305, USA.

[†]Present address: The Wistar Institute, 3601 Spruce Street, Philadelphia, PA 19104, USA.

E-mail: schung@usuhsb.usuhs.mil
subbiah@cellbio.stanford.edu

Structure 15 October 1996, 4:1123–1127

 ISSN 0969-2126

Decades of comparing and aligning protein sequences led to the empirical observation of the 'twilight zone' for sequence similarity [1]. The twilight zone is an operationally defined term. It represents a range of sequence identity that sets the boundary of confidence levels for detecting evolutionary relatedness of proteins in sequence-alignment analysis. When two protein sequences diverge, the remaining similarity, measured as percentage sequence identity, steadily decreases past the twilight zone to the limit expected by random chance. Above the twilight zone, the case for divergent evolution is strong, with greater sequence identity reflecting a shorter period of evolutionary divergence between a pair of proteins. When additional biophysical or biochemical evidence becomes available, such higher-than-twilight-zone levels of sequence identity are almost always accompanied by a very convincing similarity in three-dimensional (3D) structure and biological function. When the sequence identity falls in the twilight zone, the statistical measure for the evolutionary relatedness of proteins becomes uncertain. In most such instances, the sequences share neither an evolutionary past, similar structures nor biological functions. Despite this, there are documented cases in which seemingly unrelated sequences, sharing less than the twilight-zone limit of sequence identity, adopt similar 3D folds [2,3]. In practice, the minimum sequence identity sufficient to infer evolutionary relatedness depends on the length and amino-acid composition of the aligned sequences, as well as on the gap penalty imposed by the sequence-alignment procedures. With most computer alignment programs, the twilight zone typically falls between 20 and 25% sequence identity for proteins that are comprised of at least one stable domain [1,4]. When two totally unrelated sequences composed of the 20 standard amino acids were aligned without any introduced gaps, random chance led to a mean value of 6% for sequence identity. Sequence-alignment techniques that maximize similarity by introducing relative insertions and deletions can be expected to significantly raise this baseline average [5]. To summarize, when a pair of protein sequences have high sequence identity—higher than the twilight-zone limit of 25%—divergent evolutionary relatedness can be convincingly inferred. When the sequence identity falls within or below the twilight zone of 20–25%, common ancestry from a shared past cannot be readily assumed by sequence data alone.

Although the significance of the twilight zone is well established, much of its justification stems from our empirical experience with statistical analysis of 1D protein sequences. As almost all protein sequences fold into specific 3D structures, and these folded structures are under evolutionary selection pressure, the twilight zone is likely to have some 3D structural meaning. Cumulative amino-acid changes in protein sequence, including insertions and deletions, result in altered 3D structures [6]. The connection between sequence similarity and structural similarity can be established by a combined sequence-structural analysis of structurally superimposed proteins. Studies carried out on pairs of optimally superimposed homologous proteins demonstrated that the structural differences, measured as the average rms deviation of the backbone atoms, increases with decreasing sequence identity [7–9]. Although there are many individual exceptions, the general observation is that when two proteins share 50% or higher sequence identity, their backbones differ by less than 1 Å rms deviation; when two proteins share 20–25% sequence identity, their backbones typically differ by some 2 Å rms deviation. Therefore, to date, one simple 3D structural implication of the 1D twilight zone exists: when the sequences of two proteins diverge to the twilight-zone limit, their backbones can be expected to differ by 2 Å rms deviation. Our recent results, obtained while developing the application of side-chain-packing methods for the homology modeling of proteins, unexpectedly offers a deeper insight into the structural meaning of the twilight zone.

Approaches that analyze side-chain packing are based on the premise that the fixed backbone template of a protein is sufficient to allow the prediction of the side-chain coordinates of its buried core residues, based on packing criteria

[10]. It is now well established that different side-chain-packing methods, including the one [11] that we have used, can be expected to accurately predict the side-chain coordinates of the buried core residues when the experimental backbone coordinates of a globular protein are given [12–19]. On average, the overall side chain rms error in prediction is about 1.2 Å, while 85% of the χ_1 and 80% of the χ_2 angles can be predicted accurately [20]. Allowing for the approximate 0.3 Å experimental error in the backbone coordinates of even the most well determined X-ray structures, this is remarkably accurate when compared with the 3.1 Å side-chain rms error and the *22%* and *29%* success rates for χ_1 and χ_2 angles (averaging over all amino-acid residues) that can be expected by random chance.

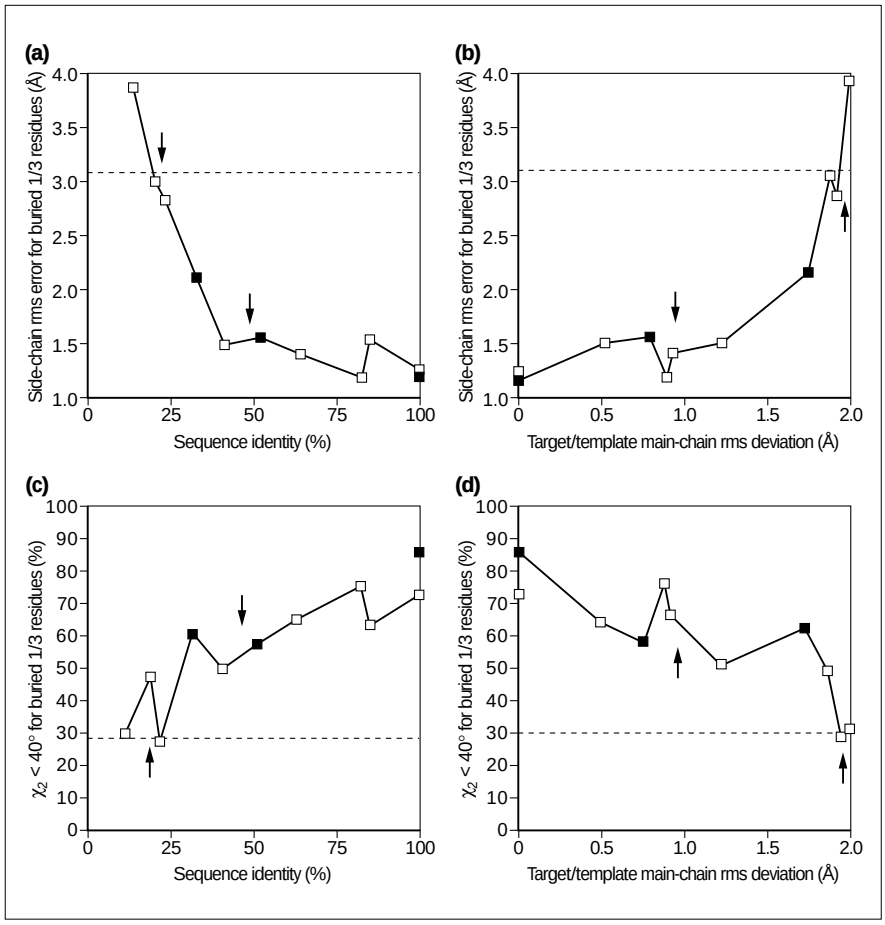
Recently, we demonstrated that the side-chain-packing methods can be successfully applied to homology modeling, using families of proteins with known 3D structures as model systems [9,21]. For each target sequence, the side-chain coordinates of the buried residues were predicted using the backbone coordinates of a known homologous protein as a fixed template. The side-chain prediction accuracy was assessed as function of either sequence similarity or backbone structural similarity between the pairs of target and template proteins (Fig. 1). We observed that the average rms errors for the predicted buried side chains increase in an exponential fashion with decreasing sequence identity or increasing backbone rms deviation (Fig. 1a,b). Specifically, when the sequence identity was about 50%, or with a corresponding backbone rms deviation of about 1 Å between the template and the true target, the average rms error for the predicted buried side chains remained low, at 1.5 Å (Fig. 1a,b, arrows). In addition, 60–65% (Fig. 1c,d, arrows) of the χ_2 angles were accurately predicted. When the template and target sequences are at the twilight zone of about 20–25% sequence identity, or the corresponding backbone rms deviation of about 1.9–2.0 Å, the prediction accuracies for the average side-chain rms error and the χ_2 angles reached their random-chance limits of 3.1 Å and 29%, respectively (Fig. 1). As the sequence identity drops from 50% to 20–25%, the χ_1 prediction accuracy, although not reaching its random limit of *22%*, is significantly reduced from 70–75 % to 50–55% [9]. The fact that χ_1 is somewhat more accurate than the expected random value is not surprising, as χ_1, unlike χ_2, is highly restricted by its own local backbone. In contrast, χ_2 is mostly restricted by its tertiary packing against the backbones of regions distal in sequence and against other nearby side chains. Like χ_2, the side-chain rms error is also mainly influenced by the degree to which the immediate tertiary environment constrains the packing possibilities of a given buried side chain. Thus, in homology modeling, when the difference between a pair of template and target proteins reaches twilight zone, the corresponding imperfect template backbone, accompanied by incompatible tertiary environment, is no longer sufficient to constrain the side chains of the target buried core residues into their correct rotamer orientations.

It is remarkable that the limit at which the template backbone is insufficient to constrain the correct packing of the buried side chains should occur at the twilight zone. This provides a structural explanation of the raison d'être for the twilight zone: the confident assumption of divergent evolution from sequence information alone. When protein sequences diverge from a common ancestor by a single amino-acid replacement in the buried core, the new side chain replacing the original residue is quite severely hemmed in by its immediate tertiary environment, comprising nearby side chains and backbone atoms distal in sequence. Thus, only certain side chains in certain rotamer conformations can be accommodated in the new folds. As more amino-acid replacements are progressively introduced, there are successive gradual distortions of the tertiary fold and side-chain interactions to accommodate the changes. However, at each step, the descendent protein's freedom to diverge is always restricted by the particular structural environment, imposed by the immediate predecessor, at the site of replacement. It appears that there is a continuing 'structural memory' of both the pattern of side-chain interactions and the constraining backbone fold endowed by the common ancestral sequence. Both these memories are embedded in the evolving repertoire of side chains, the former only in the residues that have been left unchanged and the latter in all the residues.

When the diverging protein sequences reach the twilight zone, the memory of the specific pattern of side-chain interactions endowed by the ancestral sequence is in a structural sense 'lost': major rearrangements in side-chain interactions can take place as long as the side-chain conformations are compatible with the ongoing constraining backbone scaffolds. This is consistent with our recent observation that when a pair of target and template proteins shares 20–25% sequence identity, the homologous backbone of the template structure is not sufficient to allow better than random predictions of the buried side-chain conformations [9]. In contrast to the pattern of side-chain interactions, the ancestral constraining backbone fold can, and often does, continue to be recognizable as the sequence diverges beyond the twilight zone. The well characterized examples are the globin and cytochrome c families in which pairs of homologous proteins that share less than 16% or 12% sequence identity can still be compatible with the same general constraining fold [6,22]. In the protein database, there are cases of related or unrelated proteins that share less than the twilight zone of sequence identity and yet adopt similar folds but different side-chain interactions [23,3].

The description above offers an explanation for the residual sequence identity of 20–25% seen at the twilight zone. There is a similar explanation for the corresponding

Figure 1

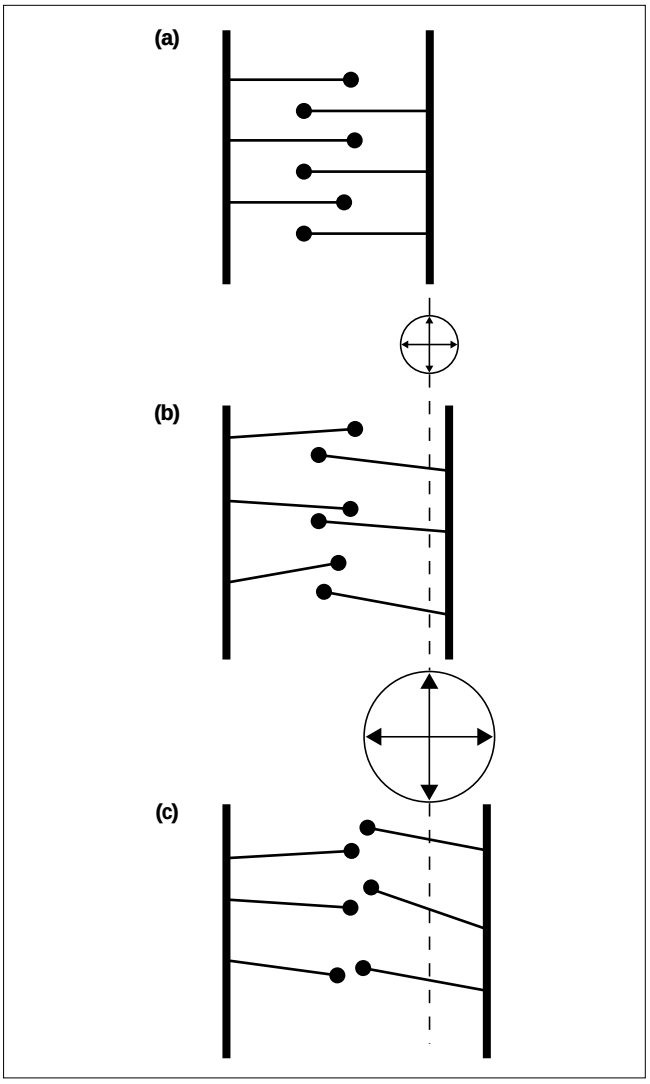


The relationship between side-chain prediction errors and sequence similarity or backbone similarity between pairs of template and target proteins. The open boxes represent globin modeling data points and the filled boxes represent the average values for data points for a bacteriophage repressor family. (a) Side-chain rms prediction error for the most well buried third of core residues against increasing sequence identity. (b) Side-chain rms prediction error against the backbone rms deviation between the homologous template structure and the target structure. (c) Percentage of successfully predicted χ_2 angles against the sequence identity. (d) The relation between the percentage of correctly predicted χ_2 angles and the backbone rms deviation between the template and target structures. The horizontal dotted line in (a) and (b) indicates the random chance limit of 3.1 Å for predicted side-chain rms error and the horizontal dotted line in (c) and (d) indicates the random chance limit of 29 % for predicted χ_2 angle accuracy. In all plots, the two arrows indicate the two regions of particular interest discussed in the text. The first arrow, at either 22 % sequence identity or at 2 Å target/template backbone rms error, corresponds to the twilight zone of protein sequence identity. The second arrow, at about 50 % sequence identity or at 1 Å target/template rms error, corresponds to an intermediate region where side-chain packing methods still give quite reliable predictions. (The figure was adapted from [9], with permission.)

1.9–2.0 Å rms backbone deviation between the pair of template and target proteins. For a backbone deviation of this magnitude, two main-chain atoms from different parts of the protein can be expected to move maximally about 3.8–4 Å closer together or further apart. As proteins are well packed in the core and must maintain such a low energy state, this corresponds to either inserting or removing an intervening spherical volume of radius 1.9–2.0 Å (i.e. an approximate volume of 33.5 Å^3). The packing volumes of the 20 amino acid side chains are well known and range from as little as 0 Å^3 for glycine to as much as 172 Å^3 for tryptophan [24]. Assuming all amino acids occur at equal frequency, the average change in volume associated with randomly selecting a residue in a protein and randomly replacing it with another can be crudely calculated to be about 35 Å^3. Such a volume change associated with random amino-acid replacement within a protein can be approximated as a relative insertion or deletion of a sphere of radius 2.0 Å. This number is very close to the 1.9–2.0 Å radius of a spherical deletion/insertion that can be maximally tolerated by a 1.9–2.0 Å rms deviation between the backbones of homologous proteins. Thus, from a purely volumetric point of view and ignoring other considerations of shape complementarity, progressive distortion of the main-chain backbone of a protein as a result of sequence divergence can reach a point where any side chain can on average be replaced by any other with little ill-effect; we find this point to correspond to the twilight zone. Stated simply, at 1.9–2.0 Å backbone rms deviation, the side chains do not 'see each other' and can behave as if they were relatively free of their constraining tertiary environment so long as the volume is sufficiently occupied and the secondary structural propensities remain more or less unchanged (Fig. 2).

In effect, this suggests a pseudo-phase transition in which specific side-chain interactions begin to be replaced by more general van der Waals' interactions. In such a scenario we envisage constant secondary structure accompanied by a fairly mobile fluidity in the orientations of the internal buried side chains and loss of specific side-chain–side-chain interactions like hydrogen bonds. This is very reminiscent of that postulated for the molten globule steps in the kinetics of the folding of proteins in solution [25]. In the molten

Figure 2



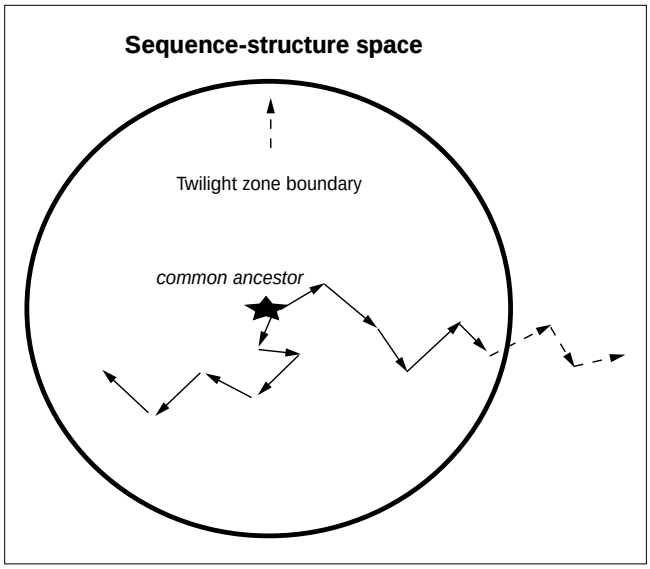
A schematic representation of the pseudo-phase transition as the backbone of a protein is progressively distorted. (a) The well packed native state with the thick vertical lines representing the connecting main-chain backbones. (Note that these do not reflect the actual backbone conformations.) The digitized horizontal pin-like lines represent the state of well-packed side chains. (b) Overall main-chain distortion at less than 2 Å rms deviation. As the main chain is gradually distorted, the side chains can wiggle a little but cannot slide past each other. The native side-chain interaction pattern is retained. The arrows in the circle indicate that main-chain atoms can shift in any direction. (c) Overall main-chain distortion for a rms deviation of greater than 2 Å. When the main-chain distortion is large enough, the side chains can pass each other. The original side-chain interaction pattern is lost and the side chains can re-pack in a different pattern. This amounts to a type of pseudo-phase transition behavior in structure space.

globule state, only the overall fold is maintained, while the specifics of the internal packing are dynamically changing. Nevertheless, such change is limited in that it has to be compatible with the overall fold. A consistent, experimental observation about the molten globule is that the volume of a protein increases in going from the fully folded native state to the molten globule state. Although experimental values are not available for the backbone rms deviation associated with these expanded protein states, computer simulations suggest a value of 2–3 Å rms deviation for the residues in the well-defined secondary structural elements that surround the stable native-like core [26]. Much of the increased volume of the so-called dry molten globule state is thought to be a result of poor packing in the relatively fluid interior [25]. At least in computer simulations, this poor packing was accompanied by a backbone distortion of some 2–3 Å rms deviation.

It is perhaps not surprising that when the target and template backbones diverge beyond a rms difference of about 2 Å, our side-chain packing methods found the template backbone to be insufficient to correctly constrain the buried side chains. Given that the twilight zone corresponds to a 1.9–2.0 Å rms backbone deviation, this implies that at the level of side-chain interactions, the native state of a protein could be about as different from its molten globule state as it is from the native state of a homologous protein that has diverged beyond the twilight zone. Thus, not unlike the constantly rearranging side chains of the molten globule, sequences that are only compatible with the backbone fold, and therefore do not necessarily share a common past, can adopt one of the limited but ancestrally unrelated repertoire of side-chain/sequence arrangements in order to pack into the overall fold.

Finally, returning to the notion of a pseudo-phase transition of the side-chain-packing pattern in structure space at the twilight zone, it would appear that an hysteresis-like behavior is associated with the twilight zone in sequence space. As illustrated in Figure 3, sequences that progressively diverge from an original common ancestor ultimately reach and cross the twilight zone boundary of 20–25% in sequence space. This is accompanied by a pseudo-phase transition in which the structural memory of specific pattern of side-chain interactions endowed by the common ancestral sequence is lost: only the ancestral fold implicit in the ongoing sequence-to-fold compatibility remains. Thus, sequences that have diverged past the twilight zone may become locked in patterns of side-chain interactions that are very different from the ancestral pattern and cannot, except by pure coincidence, evolve through evolutionary time to the original common ancestor. In contrast, sequences that have yet to cross the threshold retain the ancestral side-chain interactions and can, in principle, evolve back to the common ancestor. Esoteric as the point may be, the twilight zone constitutes a point-of-no-easy-return. Even if a sequence were to return from beyond the twilight zone toward the ancestor, the evolutionary path would probably be very different. This is reminiscent of the physical concept of hysteresis.

Figure 3



Schematic illustration of the twilight zone in sequence-structure space. The square box represents the protein sequence-structure space in which proteins evolve over time. The star in the center represents a common ancestor for a given family of divergent sequences. The circle surrounding it demarcates the twilight zone of 20–25% sequence identity. Points within the circle represent protein sequences that are unambiguously related by divergent evolution to the common ancestor and share common features of the ancestral side-chain interactions. Points outside the circle represent protein sequences that may or may not descend from the common ancestor. Each of these proteins may adopt a different pattern of side-chain interactions from that of the common ancestor. The series of arrows represent a particular trajectory of progressive mutations through evolutionary time. Within the encircled twilight zone, such trajectories (in solid arrows) can, in principle, mutate back toward the common ancestor. However, once the trajectories cross the twilight zone boundary (in dashed arrows), they will be unable, except by chance, to mutate back toward this common ancestor.

In summary, our recent work on homology modeling of target protein structures using side-chain packing methods demonstrates that the point at which a homologous template backbone is no longer sufficient to constrain the correct packing of the buried side chains occurs at the twilight zone limit of sequence identity between the template and target protein. This observation provides a 3D structural justification for the empirically-derived 1D twilight zone of sequence identity. The results suggest that protein sequences that have diverged from a common ancestor beyond the twilight zone may adopt side-chain interactions that are very different from those endowed by the ancestral sequence.

Acknowledgements

We thank Michael Levitt for providing computer programs and computing facilities, David Eisenberg and Helen Berman for helpful comments on the manuscript. This work was supported by the DOE grant DE-FG03-95ER62135 to ML and USUHS grant R071CX to SYC.

The opinions or assertions contained herein are the private ones of the authors and are not to be construed as official or reflecting the views of the Department of Defense or the Uniformed Services University of the Health Sciences.

References

1. Doolittle, R.F. (1986). *Of Urfs and Orfs: Primer on how to analyze derived amino acid sequences*. University Science Books, Mill Valley, CA, USA.
2. Subbiah, S., Laurents, D. & Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* **3**, 141–148.
3. Laurents, D., Subbiah, S. & Levitt, M. (1994). Different protein sequences can give rise to highly similar folds through different stabilizing interactions. *Protein Sci.* **3**, 1938–1944.
4. Sander, C. & Schneider, R. (1991). Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**, 56–68.
5. Doolittle, R.F. (1981). Similar amino acid sequences: chance or common ancestry? *Science* **214**, 149–159.
6. Lesk, A.M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225–270.
7. Chothia, C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
8. Flores, T.P., Orengo, C.A., Moss, D.S., & Thornton, J.M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **3**, 2358–2365.
9. Chung, S.Y. & Subbiah, S. (1996). How similar must a template protein be for homology modeling by side-chain packing methods. In *Proceedings of the first Pacific Symposium on Biocomputing*. (Hunter, L., & Klein, T., eds), pp. 126–141, Hawaii, USA.
10. Ponder, J.W. & Richards, F.M. (1987). Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791.
11. Lee, C. & Subbiah, S.J. (1991). Prediction of protein side-chain conformation by packing optimization. *Mol. Biol.* **217**, 373–388.
12. Holm, L. & Sander, C.J. (1991). Database algorithm for generating protein backbone and side-chain coordinates from a Cα trace application to model building and detection of coordinate errors. *Mol. Biol.* **218**, 183–194.
13. Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R.J. (1991). A new approach to the rapid determination of protein side chain conformations. *Biomolec. Struct. Dyn.* **8**, 1267–1289.
14. Desmet, J., De Maeyer, M., Hazes, B., & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain prediction. *Nature* **356**, 539–542.
15. Dunbrack Jr., R.L. & Karplus, M.J. (1993).Backbone-dependent rotamer library for proteins: application to side-chain prediction. *Mol. Biol.* **230**, 543–574.
16. Eisenmenger, F., Argos, P. & Abagyan, R. J. (1993). A method to configure protein side-chain from the main-chain trace in homology modeling. *Mol. Biol.* **231**, 849–860.
17. Wilson, C., Gregoret, L.M. & Agard, D.A. (1993). Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.* **229**, 996–1006.
18. Koehl, P. & Delarue, M.J. (1994). Application of a self-consistent mean field theory to predict protein side-chain conformation and estimate their conformational energy. *Mol. Biol.* **239**, 249–275.
19. Lee, C.J. (1994). Prediction of protein mutant energetics by self-consistent ensemble optimization. *Mol. Biol.* **236**, 918–939.
20. Tanimura, R., Kidera, A. & Nakamura, H. (1994). Prediction of protein mutant energetics by self-consistent ensemble optimization. *Protein Sci.* **3**, 2358–2365.
21. Chung, S.Y. & Subbiah S. (1995). The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins. *Protein Sci.* **4**, 2300–2309.
22. Creighton, T.E. (1993). *Proteins: structures and molecular properties*. Freeman, NY, USA.
23. Russell, R.B. & Barton, G.J. (1994). Structural features can be unconserved in proteins with similar folds. *J. Mol. Biol.* **244**, 332–350.
24. Chothia, C. (1975). Structural invariants in protein folding. *Nature* **245**, 304–308.
25. Ptitsyn, O. (1992). The molten globule state. In *Protein Folding*. (Creighton, T.E., ed.), pp. 243–300, Freeman, NY, USA.
26. Daggett, V. & Levitt. M. (1992). A model of the molten globule state from molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **89**, 5142–5146.