A PROTEIN CONFORMATIONAL SEARCH SPACE DEFINED BY SECONDARY STRUCTURE CONTACTS

M. PARISIEN, F. MAJOR
Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal,
Montréal, Québec, Canada H3C 3J7
{parisien|major}@iro.umontreal.ca

M. PEITSCH
Glaxo Wellcome Research & Development and
Geneva Biomedical Research Institute,
CH-1228 Plan-les-Ouates, Switzerland
mcp13936@ggr.co.uk

A conformational search space describing the relative position and orientation of protein secondary structure elements in three-dimensions was defined. These spatial relations were encoded by homogeneous transformation matrices between pairs of residues "in contact" in two different secondary structure elements. A database of all occurrences of spatial relations for five hydrophobic residues was built. The use of one residue contact per pair of secondary structure elements, which were approximated by standard (ϕ, ψ) assignments, was sufficient to reproduce accurately the core structure of proteins with known three-dimensional structures.

1 Introduction

Knowledge about protein three-dimensional (3-D) structure is key to protein function comprehension and manipulation. Due to difficulties associated with experimental protein structure elucidation, it is not surprising that predictive methods are increasingly gaining popularity. Protein modeling is mainly restricted to comparative methods which only apply to 15 to 20 percent of all known sequences sharing more than 30% identity.[1–5] Consequently to the many genome sequencing projects, an explosion of novel gene discoveries of unknown structure and function is observed.[6] *De novo* protein structure prediction methods are thus needed.

The secondary structure (SS) of a protein can be inferred from its sequence by using statistical methods, such as Markov models[7,8] and neural networks.[9,10] The SS of a protein can also be determined experimentally, for instance from NMR spectroscopy data. The β-sheet topology of a protein can also be inferred from statistical methods,[11] and determined from NMR spectroscopy data.[12,13] Once the sheet topology has been assigned, atomic coordinates of homologous β-sheets in previously determined 3-D structures

can be proposed. Thus, α-α and α-β residue contacts can be inferred theoretically or determined experimentally.[14,15] These contacts can be translated into geometrical constraints to define a *constraint satisfaction problem* (CSP) to resolve the 3-D structure.

In the search for an acceptable *de novo* modeling scheme, existing methods have been considered and analyzed, and our desire to make use of accumulated structural data led us to consider a protein adaptation of the MC-SYM RNA CSP solver.[16–18] We thus propose the following scheme for *de novo* protein structure prediction: (i) the definition of the protein SS by existing experimental and theoretical methods; (ii) the use of SS information to assign β-sheet topologies and α-α and α-β residue contacts to define a CSP; (iii) the use of MC-SYM to generate consistent core structures; (iv) the use of existing methods to complete the core structures with loops and side chains; and, (v) the refinement and evaluation of the structures using existing energy minimization protocols and potentials.[19–21]

In this article, the focus has been put on the implementation of the protein conformational search space, the creation of operators to manipulate protein 3-D core structures, and a best-first search algorithm to demonstrate that the developed conformational search space contains the native x-ray crystal structures. Other conformational search spaces were introduced in the past. The most common methods are based on the sampling of the ϕ-ψ torsion space,[22] on theoretical spatial relations of SS elements,[14,23] on the properties of the loops connecting the SS elements,[21] and on geometrical samplings of the SS element space.[24] Although almost the same precision can be reached by the use of these methods, they describe conformational search spaces that are larger than the one introduced in this article, and, in general, require more structural information to converge to the native fold.

2 Conformational search space

2.1 Definitions

Residue contacts bears side-chain and backbone packing information, that is, the relative position and orientation of the two SS elements which contain the residues in contact. A protein *core structure* is the assembly of its constituent SS elements in 3-D space. Two SS elements are in *SS contact* if they share at least one residue contact.

A *residue contact* between residues A and B forms if their distance is smaller than a certain threshold, $|A, B| < d$, where d is the threshold value and $|\bullet|$ denotes the Euclidean distance. The ensemble of all residue contacts in

a given protein constitutes a *residue contact graph*, where each node represents a residue and each edge represents a residue contact (see Figure 1).

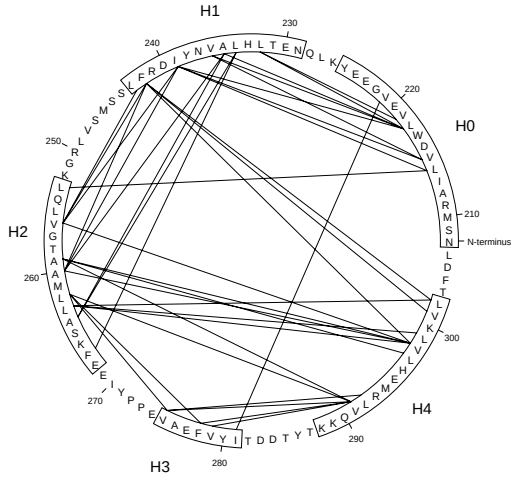

Figure 1: Residue contact graph for the cyclin box (PDB file 1fin). The framed residues are those involved in SS elements. The lines represent residue contacts. The distance threshold was set to 7.0Å between threading points as defined in reference 19.

Similarly, the ensemble of all SS contacts defines the *SS contact graph*, where the nodes represent the SS elements and the edges indicate that at least one residue contact exists between a given pair of residues in the connected SS elements (see Figure 2).

Every SS element in a protein is involved in a SS contact. To satisfy this condition, consider the degenerated distance threshold, $d = \infty$. This makes the SS contact graph *connected*, that is, there is a *path* that connects any pairs of SS elements. A connected SS contact graph that contains no cycle is a *SS contact spanning tree* (see Figure 3). There are N^{N-2} spanning trees for a graph that contains N fully connected vertices. The SS spanning tree addresses all SS elements and suggests a construction order in which the SS elements can be introduced. A possible order for the SS spanning tree in Figure 3 would be: H2 as the reference SS element; H1 placed from H2; H3 placed from H2; H4 placed from H3; and, H0 placed from H3. It is thus possible to define a protein *conformational search space* from a SS contact graph. Each residue contact can either be used as a spatial relation operation which positions and orients

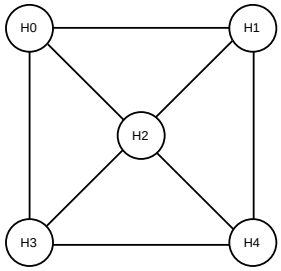

Figure 2: SS contact graph for the cyclin box (PDB file 1fin). The SS elements are circled. An edge was drawn when at least one residue contact was observed between two SS elements (see the residue contact graph in Figure 1).

a SS element from another one (just as in the construction order above), or as a distance constraint. For instance, the contacts dropped in the selection of the SS spanning tree should be replaced by distance constraints that must be satisfied in the final constructions.

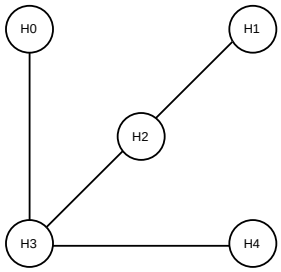

Figure 3: One of the spanning trees for the cyclin box (PDB file 1fin). The SS elements are circled. An edge was drawn when at least one residue contact was observed between two SS elements. The tree, as compared to the graph contains no cycle (see the corresponding SS contact graph in Figure 2.)

2.2 Implementation

Homogeneous transformation matrices[25] (HTM) were used to encode the spatial information of residue contacts. HTMs contain translation and rotation information. For instance, the *local referential* of a residue can be represented by an HTM from three right-handed unary orthogonal vectors that can be calculated from three non-colinear atomic coordinates. A local referential indicates

the translation and rotation to be applied to the residue coordinates expressed in the canonical referential to obtain its absolute coordinates. Consider for instance the local referential of a residue A, R_A, which can be calculated by using three backbone atoms in A. One of the three is elected as the origin of A while the two others respectively align with the X and Y axes (see Figure 4). Backbone atoms, instead of side-chain, were chosen because the backbone characterizes much better the relative orientation and position of SS elements.

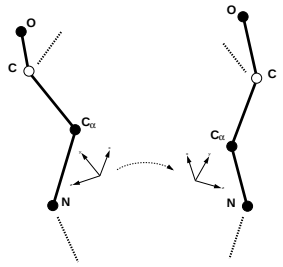

Figure 4: Spatial relation between two residues. The axis systems represent the local referential of the residues. The dotted arrow indicates the transformation of one's referential into the other. The atoms selected to compute the local referential are indicated with black dots. The dotted lines indicate the peptide bonds.

The spatial relation between residues A and B can also be expressed with an HTM: $T_{A\to B} = R_A^{-1} \times R_B$. A residue contact between A and B can be reproduced between any pair of residues, for instance A' and B', by applying $R_{B'}^{-1} \times T_{A\to B} \times R_{A'}$ to the atomic coordinates of B' to position and orient B' with respect to A'. Symmetrically, $R_{A'}^{-1} \times T_{A\to B}^{-1} \times R_{B'}$ applied to the atomic coordinates of A' positions and orients A' relative to B'.

Any residue contact found in the Protein Data Bank (PDB)[26] can be extracted and used afterwards as a building block of protein 3-D structure to orient and position SS elements. Once a pair of residues have been positioned and oriented from the application of HTMs, the extension of each SS element is made by using standard (ϕ, ψ) assignments for the other residues; for instance, $(-60°,-40°)$ for α-helices and $(-120°, 140°)$ for β-strands.[27] In this way, any pair of SS elements involved in SS contact in the database of known 3-D structures can accurately be reproduced from a single residue contact. A protein 3-D structure can be built by applying this construction scheme to each of its constituent SS elements. Our hypothesis is that all protein 3-D folds are contained in a conformational search space defined from such SS element contacts.

2.3 Transformational sets

A *transformational set* is a set of HTMs associated with a residue contact type, that is, the types of residues and the nature of their host SS elements (α-α, α-β, intra-strand β-β and inter-strand β-β). All possible residue combinations could be part of a residue contact, and thus $1600 = 4 \times 20 \times 20$ transformational sets could be defined. A question that was addressed is whether it would be possible to find a smaller subset of residue contacts that would allow one to define a conformational space containing all protein folds within a desired precision.

Subsets of residues can be identified from *weighted* SS contact graphs where the *weight* of an edge is determined by the minimum *residue contact distance* between the connected SS elements (see Figure 5). The residue contact distance is defined by the Euclidean distance between the two closest threading points of two residues, as defined in reference 19. If only a subset of residues is considered then a subset-specific weighted SS contact graph is defined. The minimum SS spanning tree can be computed from this graph, as shown in Figure 6. The number of contacts in the spanning tree depends on the relative frequencies of the residues and their propensity to make contact. The magnitude in the distances is function of SS element packing and the nature of the contacts.

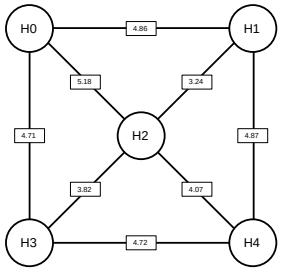

Figure 5: Weighted SS contact graph for cyclin box (PDB 1fin). The weights correspond to the minimum distances between two connected SS elements.

Consider, for instance, all subsets of five residues[a]. There are $\frac{20!}{5!(20-5)!} = 15504$ such subsets. For each subset, consider all contacts and contact distances found in the minimum SS spanning trees obtained from all protein 3-D structures in the PDB Select 25 database (the main characteristic of the PDB Select 25 is that no two structures share more than 25% sequence identity[b]).[28,29] The

[a] the number five came from an initial intuition that the subset composed of the five most hydrophobic residues could generate good results.

[b] note that a database containing no similar folds would be more appropriate.

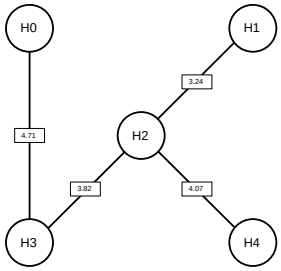

Figure 6: Weighted SS spanning tree for cyclin box (PDB 1fin). This is the minimum spanning tree corresponding to the graph in Figure 5.

subset that maximizes the number of contacts and returns the smallest mean and median distances is {ALA, ILE, LEU, PHE, VAL}. This result is somewhat not surprising since hydrophobic residues are known to be buried inside proteins and form contacts. This result also confirms our intuition that hydrophobic residues could be best suited for the proposed construction scheme.

A database of transformational sets was thus built using the subset {ALA, ILE, LEU, PHE, VAL} for α-α, α-β and inter-strand β-β contacts defined by distances smaller than 7.0Å. For the intra-strand $\beta - \beta$ contacts, hydrogen bonds were used.

3 Demonstration

Here we demonstrate that the conformational search space defined by the transformational sets defined above contains any of the known x-ray crystal structures. The demonstration is based on the reproduction of the x-ray crystal structures of 46 proteins randomly taken from PDB-Select 25. Note that the reproductions of those proteins were made by using HTMs extracted from other proteins (Jack Knife experiment[30]). To avoid exhaustive exploration, the building procedure was driven by knowledge of the x-ray crystal structure and, in the consideration of several possible search directions, by exploring first the ones minimizing the root mean square deviation (RMSD) from the x-ray crystal structure; for this reason, the procedure is referred to as the best-first search procedure. This algorithm, however, does not guarantee convergence to the optimal construction.

1. Build a pseudo-crystal structure. Make the reference element the one of the two SS elements that are linked by the largest number of residue contacts and add it to the best-first queue.

2. If the queue is empty then STOP and report the best conformation found so far. Otherwise, select from the best-first queue the partial structure that minimizes the RMSD from the crystal structure divided by the number of SS elements in the partial structure, and superimpose it to the crystal structure.

3. Append to the partial structure selected in step 2 a new SS element according to the spanning tree. All residue contacts from the crystal structure can be used to append the new element. For each contact, compute the spatial relation between the partial structure residue and its partner in the pseudo-crystal structure, T_{best}. Among the transformational set for this contact, determine the best HTM by taking the matrix, from those that differ from T_{best} by less than δÅ in the translation, and that minimizes the Euclidean distance of the rotations. Apply the best HTM and the canonical ϕ-ψ assignments to position and orient the SS element in the partial structure. Add the new partial structure to the best-first queue.

4. If the new partial structure from step 3 is complete, that is, all SS elements are present, then compare it to the best completed structure built so far and select the one that has the minimum RMSD with respect to the pseudo-crystal. Remove from the best-first queue all partial structures that would lead to higher RMSDs. A lower bound for the RMSD of partial structures is approximated by adding 0.15Å for each missing SS element. Thus, partial structures with a lower bound RMSD higher than the current best RMSD are eliminated from the queue. Goto step 2.

In the first step of this algorithm, a pseudo-crystal structure is built. The pseudo-crystal represents the x-ray crystal structure from which SS elements were substituted by standard ϕ-ψ assignment SS elements. The building order that was considered is a maximum spanning tree derived from a weighted graph where the nodes represent SS elements and the vertices represent residue contacts (see Figures 7 and 8). The weights were defined as the number of residue contacts observed in the x-ray crystal structure.

The results of applying the best-first search procedure to 46 proteins are shown in Table 1. The results suggest that spatial relations among the five selected hydrophobic residues "in contact" can be used as building blocks of protein 3-D structures. From the RMSD values, the x-ray crystal structures of all tested proteins are clearly accessible from a conformational search space defined by residue contacts.

Table 1: Results of the best-first search for 46 proteins of the PDB. Proteins are referred to by their PDB mnemonics. The RMSD^1 values indicate the RMSD of the pseudo-crystal structure from the crystal structure. The pseudo-crystal is obtained by substituting the SS elements by canonical elements obtained from standard assignments for the α-helices and the β-strands. The RMSD^2 values indicate the RMSD of the best found structure, as identified by the best-first search procedure, from the corresponding pseudo-crystal structures. The RMSD^3 values indicate the RMSD of the best found structure, as identified by the best-first search procedure, from the corresponding x-ray crystal structures. The #PU values indicate from how many different proteins the HTMs used in the best structure were extracted. Note that for proteins composed of N SS elements, $N-1$ HTMs were used for its construction. All RMSD values were calculated for the backbone atoms in the SS elements only.

Protein	#SS		#residues		#PU	RMSD^1 (Å)	RMSD^2 (Å)	RMSD^3 (Å)
	α	β	α	β				
1aak	5	4	44	26	7	0.83	2.47	3.09
1ab2	2	8	20	37	8	1.27	2.88	3.29
1abm	7	5	110	24	11	1.45	3.33	3.26
1atx	0	4	0	22	3	1.52	3.98	4.40
1bab	7	0	108	0	6	0.73	2.54	2.60
1bvh	5	4	52	20	8	0.79	2.44	2.36
1c5a	4	0	49	0	3	0.79	1.56	1.95
1cbn	2	2	21	8	3	0.66	2.17	1.62
1cde	6	7	88	52	12	1.28	2.88	2.85
1chr	12	11	135	57	21	1.04	2.92	3.08
1crl	11	11	116	65	20	1.32	3.26	3.20
1dsb	8	5	112	30	12	1.07	2.97	3.07
1ede	11	8	121	46	18	1.19	3.54	4.26
1erg	1	3	8	14	3	1.15	2.10	2.03
1fas	0	3	0	18	2	1.29	2.07	2.88
1gox	11	8	117	32	18	0.71	2.53	2.77
1hlb	8	0	115	0	7	0.86	2.24	2.39
1l18	9	4	102	15	12	0.69	2.93	2.78
1lga	13	0	125	0	11	0.91	2.49	2.95
1nar	7	9	87	55	14	0.96	2.81	2.92
1ofv	5	6	53	27	10	0.75	2.31	2.60
1pfk	13	11	168	66	21	0.95	2.73	3.06
1phh	12	18	125	81	26	0.86	3.42	3.18
1pii	18	16	157	79	26	0.62	3.35	3.20
1pox	19	20	229	113	32	0.87	3.52	3.59
1rhd	10	10	109	43	19	0.93	3.57	3.38
1sbp	12	11	131	57	20	1.00	3.15	3.62
1sto	7	5	91	34	11	1.75	3.11	3.15
1tca	10	7	89	35	16	1.67	3.04	3.54
1tml	7	7	86	35	12	0.73	2.39	2.38
1ula	7	12	87	68	18	1.08	3.12	3.06
1wsy	11	8	117	50	18	0.78	2.77	3.40
1xya	13	9	170	51	18	0.81	2.83	2.66
2acq	10	8	112	41	16	0.71	3.34	3.22
2atc	9	11	119	81	10	1.41	3.18	3.37
2ctc	9	8	110	45	16	0.74	2.76	2.91
2cyp	12	0	149	0	9	0.99	2.59	2.46
2fal	8	0	112	0	7	0.88	2.16	2.22
2gbp	10	12	139	67	17	1.07	2.64	3.17
2liv	4	7	54	44	9	1.20	2.51	2.53
2pia	5	11	41	70	13	2.55	3.60	3.52
2rn2	4	5	47	43	8	0.91	2.61	2.63
2tmd	16	13	162	69	26	1.00	3.89	3.91
3dfr	3	8	33	59	10	1.58	2.61	3.06
4fxn	4	5	52	37	8	1.33	2.99	3.00
5p21	4	6	51	44	9	1.31	2.86	3.24

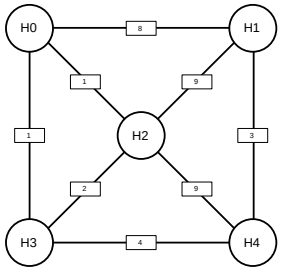

Figure 7: Weighted SS contact graph for cyclin box (PDB file 1fin). The weights in this case correspond to the number of residue contacts between two connected SS elements.

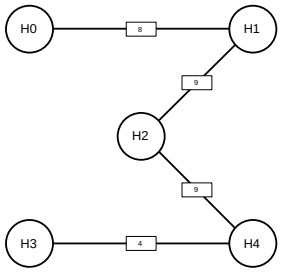

Figure 8: Weighted SS spanning tree for cyclin box (PDB file 1fin). This is the maximum spanning tree corresponding to the graph in Figure 7.

4 Conclusion

A new and efficient representation of protein conformational search space, based on residue contacts, was developed. We have shown that hydrophobic contacts contain information about the relative position and orientation of α-α and α-β SS elements. This is of course only a necessary step, not a highly significant result, since evaluating and deciding which fold is the correct one represents the actual difficulty of automated protein structure determination. Nevertheless, the technique presented here shows promises for the development of a productive protein 3-D modeling scheme. For instance, the technique should allow one to explore a representative small fraction of a protein's conformational space with the use of low resolution data, such as covariation data from multiple sequence analysis and mutagenesis data. Furthermore, a better characterization of residue contacts and of their spatial relations should allow us to predict protein 3-D structure from sequence and SS information, a

requisite to *de novo* protein design. The fact that higher RMSD values were measured for proteins that are mainly composed of β-strands indicates that more efforts should be put on the re-construction of β-sheets than on the re-construction of α-helices. Producing actual predictions is the next step of this research project.

Acknowledgments

We thank S. Lemieux and S. Oldziej for interesting discussions about this work. This work is funded by the Medical Research Council (MRC) of Canada. MP is a NSERC graduate scholar. FM is a MRC fellow.

References

1. R.F. Doolittle, *Science* **214**, 149 (1981).
2. P. Bork *et al*, *Nature* **358**, 287 (1992).
3. C. Chothia, *Nature* **357**, 543 (1992).
4. G. Casari *et al*, *Nature*, **376**, 647 (1995).
5. J. Moult, *Curr. Opin. Biotech.* **7**, 422 (1996).
6. NCBI-Genbank (*http://www.ncbi.nlm.nih.gov/*, 1997).
7. K. Asai, S. Hayamizu, and K. Handa, *CABIOS* **9**, 141 (1993).
8. A. Krogh *et al*, *J. Mol. Biol.* **235**, 1501 (1994).
9. N. Quian and T.J. Sejnowski, *J. Mol. Biol.* **202**, 865 (1988).
10. L.H. Holley and M. Karplus, *Proc. Natl. Acad. Sci. (USA)* **86**, 152 (1989).
11. T.J.P. Hubbard in *Proc. Biotech. Comp. Track (HICSS)*, ed. R.H. Lathrop (IEEE Computer Society Press, 1994).
12. W. Zhang, T.E. Smithgall and W.H. Gmeiner, *FEBS Letters* **406**, 131 (1997).
13. H. Ponstingl and G. Otting, *Eur. J. Biochem.* **244**, 384 (1997).
14. F.E. Cohen, M.J.E. Sternberg, and W.R. Taylor, *J. Mol. Biol.* **156**, 821 (1982).
15. K.-C. Chou *et al*, *J. Mol. Biol.* **186**, 591 (1985).
16. F. Major *et al*, *Science* **253**, 1255 (1991).
17. D. Gautheret, F. Major, and R. Cedergren, *J. Mol. Biol.* **229**, 1049 (1993).
18. F. Major, D. Gautheret, and R. Cedergren, *Proc. Natl. Acad. Sci. (USA)* **90**, 9408 (1993).
19. Y. Kaizhi and K.A. Dill, *Protein Science* **5**, 254 (1996).
20. S.H. Bryant and C.E. Lawrence, *Proteins* **16**, 92 (1993).

21. J. Gunn, *J. Chem. Phys.* **106**, 4270 (1997).
22. M.J. Rooman, J.-P.A. Kocher and S.J. Wodak, *J. Mol. Biol.* **221**, 961 (1991).
23. T.R. Defay and F.E. Cohen in *Encyclopedia of Molecular Biology and Molecular Medicine*, ed. R.A. Meyers (VCH Publishers Inc., New-York, NY, 1996).
24. P. Herzyk and R.E. Hubbard, *Biophys. J.* **69**, 2419 (1995).
25. R.P. Paul, *Robot Manipulators: Mathematics, Programming, and Control* (MIT Press, Cambridge, MA, 1981).
26. F.C. Bernstein *et al*, *Eur. J. Biochem.* **80**, 319 (1977).
27. J.S. Richardson, *Adv. Prot. Chem.* **34**, 167 (1981).
28. U. Hobohm *et al*, *Protein Science* **1**, 409 (1992).
29. U. Hobohm and C. Sander, *Protein Science* **3**, 522 (1994).
30. B. Efron, *The Jack Knife, the Bootstrap and Other Resampling Plans* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982).