Journal of Mathematical Chemistry 8(1991)217–227

AN ALGORITHM FOR CONVERTING A VIRTUAL-BOND CHAIN INTO A COMPLETE POLYPEPTIDE BACKBONE CHAIN

Ning LUO, Masayuki SHIBATA and Robert REIN
Department of Biophysics, Roswell Park Memorial Institute, Buffalo, NY 14263, USA

Abstract

A systematic analysis is presented of the algorithm for converting a virtual-bond chain, defined by the coordinates of the α-carbons of a given protein, into a complete polypeptide backbone. An alternative algorithm, based upon the same set of geometric parameters used in the Purisima – Scheraga algorithm [1] but with a different "linkage map" of the algorithmic procedures, is proposed. The global virtual-bond chain geometric constraints are more easily separable from the loal peptide geometric and energetic constraints derived from, for example, the Ramachandran criterion [2, 3], within the framework of this approach.

1. Introduction

The problem of protein folding, i.e. the prediction of the three-dimensional structure of a protein from its primary sequence of amino acids, remains one of the most difficult problems in biophysical chemistry. The origin of the difficulty lies in the fact that protein folding is a typical "NP (Nondeterministic Polynomial)-complete" problem [4]. In other words, the complexity of the problem increases exponentially with the number of sites in the problem. For a polypeptide sequence of 100 residues, each with ten possible configurations, there are $\sim 10^{100}$ configurations to be examined and tested when searching for a stable folded structure. There are basically two ways to tackle an NP-complete problem. One is to reduce the complexity into a "P-complete" problem by constructing some models with less heterogeneity in them [5, 6]. However, since structural heterogeneity is essential for the biological specificity of an enzymatic protein, any attempt to reduce its complexity to a "P-complete" degree by a homogeneous modeling will inevitably lose most of its essential features. The other approach to the NP-complete problem is to seek aid from empirical data so that the growth in complexity, although still NP-complete, is under the control of the investigator. For example, one of the approaches to the static structure of a folded protein is the so-called virtual-bond method [7], in which a complete polypeptide backbone chain is represented by a given set of C_i^{α} coordinates (i is the residue number) and the "virtual bonds" connecting them.

To compare the results of the theoretical modeling by the virtual-bond method with experimental data, one needs to find all the backbone conformations compatible

with a given virtual chain. Purisima and Scheraga designed an algorithm in 1984 [1] to convert a virtual-bond chain into a complete main chain based on the following assumptions:

(i) all bond lengths and bond angles are fixed; and

(ii) all peptide bonds are in the "*trans*" conformation, i.e. $\omega_i = 180° \; \forall i$.

The Purisima–Scheraga algorithm (PS algorithm) has been tested by the Scheraga group [1] and in this laboratory (see, e.g. [8,9]) on various proteins.

In addition, algorithms for generating full backbone conformations from a given set of C_i^α coordinates are also very useful in studying a protein's structure when only its C_i^α coordinates are available from X-ray crystallography.

In this article, a systematic analysis of the algorithms for generating polypeptide main chain conformations under the given geometric constraints of the virtual chain is carried out. An alternative to the PS algorithm that is based on the same set of the assumptions and geometric parameters has emerged. The geometric set-up and the notational definitions of the problem are briefly reviewed in section 2. In section 3, the "linkage maps", i.e. the computational procedural relations among the variables, for the PS algorithm are presented. In section 4, the new algorithm is constructed, based on an analysis of the role played by the auxiliary parameters $(\lambda_1)_i$ and $(\lambda_2)_i$ introduced in the PS algorithm. Finally, the similarities and the differences between the two algorithms and the possible advantages of the new algorithm are summarized in section 5.

2. Geometry and notations of the virtual-bond method

Following the standard convention [10], ECEPP geometry [11] and notations of Purisima and Scheraga [1], the geometric parameters involving the virtual-bond chain and the polypeptide main chain are (see fig. 1 [12]):

α_i = bond angle formed by $N_i C_i^\alpha C_i'$ (from ECEPP geometry),

ξ_i = angle formed by $N_i C_i^\alpha C_{i-1}^\alpha$ (from ECEPP geometry),

η_i = angle formed by $C_i' C_i^\alpha C_{i+1}^\alpha$ (from ECEPP geometry),

ζ_i = virtual-bond angle formed by $C_{i-1}^\alpha C_i^\alpha C_{i+1}^\alpha$,

γ_i = virtual dihedral angle formed by $C_{i-1}^\alpha C_i^\alpha C_{i+1}^\alpha C_{i+2}^\alpha$ (fig. 2),

ϕ_i = dihedral angle formed by $C_{i-1}' N_i C_i^\alpha C_i'$,

ψ_i = dihedral angle formed by $N_i C_i^\alpha C_i' N_{i+1}$,

$(\lambda_1)_i$ and $(\lambda_2)_i$ = auxiliary dihedral angles that define the orientation of the two planar peptide groups adjacent to C_i^α with respect to the plane defined by the virtual-bond angle $C_{i-1}^\alpha C_i^\alpha C_{i+1}^\alpha$.

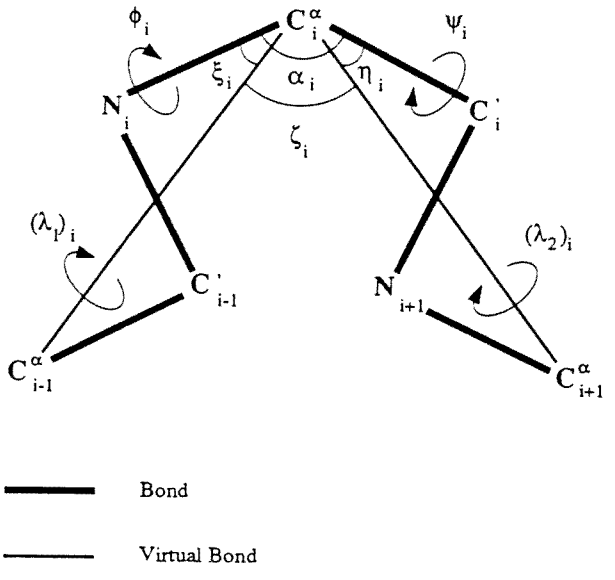


Fig. 1. The definitions of the angular parameters α, ζ, ξ, η, ϕ, ψ, λ_1 and λ_2 (see ref. [12]).

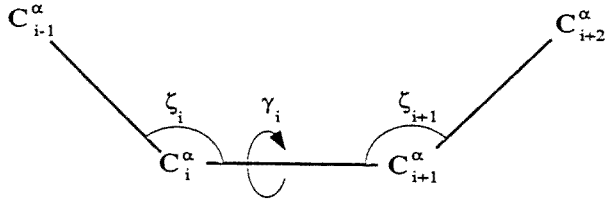


Fig. 2. The definitions of the angular parameters in the virtual-bond chain structure (see ref. [12]).

Note that α_i, ξ_i, η_i, ζ_i and γ_i are constant parameters fixed either by the defined geometry or by the given virtual-bond structures, while $(\lambda_1)_i$, $(\lambda_2)_i$, ϕ_i and ψ_i are the variables to be computed in the algorithm.

There are two basic relations, derived by Nishikawa et al. [12] from the geometry defined in fig. 1. First,

$$\mathbf{T}^x_{\lambda_1}\mathbf{T}^z_{(\pi-\zeta)}\mathbf{T}^x_{\lambda_2} = \mathbf{T}^z_{\xi}\mathbf{T}^x_{\phi}\mathbf{T}^z_{(\pi-\alpha)}\mathbf{T}^x_{\psi}\mathbf{T}^z_{\eta}, \tag{1}$$

where $\mathbf{T}^x_\theta$ (or $\mathbf{T}^z_\theta$) is the rotation matrix of angle θ (where θ represents any of the angles appearing in eq. (1)) with respect to the x-axis (or z-axis) and all angles refer

to the same residue i. Second, from the definitions of $(\lambda_1)_i$ and $(\lambda_2)_i$ in fig. 1 and from fig. 2:

$$\gamma_i = (\lambda_2)_i + (\lambda_1)_{i+1} + \pi, \tag{2}$$

which links the variable $(\lambda_2)_i$ of one residue to the variable $(\lambda_1)_{i+1}$ of the next residue.

Since all the **T** matrices in eq. (1) are rotational matrices, each of them or any product of them is an orthogonal matrix. There are three independent components in a 3×3 orthogonal matrix, which can be specified by, for example, the three independent Euler angles. Therefore, once one of the four variables $(\lambda_1, \lambda_2, \phi, \psi)$ of a residue is determined, the other three variables can be obtained by solving the component equations in eq. (1). Such solutions may not be unique, though, since the orthogonality constraint for each angle θ

$$\sin^2\theta + \cos^2\theta = 1, \tag{3}$$

i.e. the Pythagorean theorem, is nonlinear.

It is not necessary to put all variable angles under the constraint of eq. (3). In fact, only one angle (ϕ in backward generation and ψ in forward generation) is constrained by eq. (3) in the PS algorithm so that the sine of the angle is not independent of the cosine of the angle. For all other unknown pairs of angles, the sines and the cosines are treated as independent quantities. The PS algorithm uses five equations derived from eq. (1) and one equation of the type eq. (3) to solve for three variable angles' sines and cosines, based on the values of $\cos\lambda_l$ and $\sin\lambda_l$ (where $l = 1$ or 2) derived from the neighboring residue by eq. (2).

3. A graphical analysis of the matrix equation and the algorithmic procedures

Since the choice of the set of equations needed to solve the variable angles is not unique, it is interesting to ask: Could the search of the full backbone chain be improved by choosing a different set of equations? In order to satisfy this problem in a more systematic way, a concept of "contraction" and a graphic representation of deriving algebraic equations from the matrix equation (1) are introduced in the following discussion.

Because all of the matrices in eq. (1) are rotation matrices with respect to either the x-axis or the z-axis, sandwiching any product of the matrices with the eigenvectors

$$\boldsymbol{u}_x \equiv \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{or} \quad \boldsymbol{u}_z \equiv \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

and their transposes $\boldsymbol{u}_x^\dagger$ and $\boldsymbol{u}_z^\dagger$ will eliminate two of the T matrices together with the two associated parameters from the resulting expression. For example, sandwiching eq. (1) with $\boldsymbol{u}_x^\dagger$ and $\boldsymbol{u}_x$ gives

$$\boldsymbol{u}_x^\dagger \mathbf{T}^z_{(\pi-\zeta)} \boldsymbol{u}_x = \boldsymbol{u}_x^\dagger \mathbf{T}^z_\xi \mathbf{T}^x_\phi \mathbf{T}^z_{(\pi-\alpha)} \mathbf{T}^x_\psi \mathbf{T}^z_\eta \boldsymbol{u}_x , \tag{4}$$

in which two variables λ_1 and λ_2 no longer appear. This is in fact one of the equations used in the PS algorithm. Its left-hand side is

$$\boldsymbol{u}_x^\dagger \mathbf{T}^x_{\lambda_1} \mathbf{T}^z_{(\pi-\zeta)} \mathbf{T}^x_{\lambda_2} \boldsymbol{u}_x = \boldsymbol{u}_x^\dagger \mathbf{T}^z_{(\pi-\zeta)} \boldsymbol{u}_x = -\cos\zeta, \tag{5}$$

while the right-hand side is

$$\boldsymbol{u}_x^\dagger \mathbf{T}^z_\xi \mathbf{T}^x_\phi \mathbf{T}^z_{(\pi-\alpha)} \mathbf{T}^x_\psi \mathbf{T}^z_\eta \boldsymbol{u}_x = a\cos\xi + b\sin\xi\cos\psi - \sin\xi\sin\phi\sin\psi\sin\eta, \tag{6}$$

where

$$a \equiv \cos\alpha\cos\eta + \sin\alpha\cos\psi\sin\eta, \tag{7}$$

$$b \equiv \sin\alpha\cos\eta - \cos\alpha\cos\psi\cos\eta. \tag{8}$$

If eq. (1) is rewritten as:

$$I = \mathbf{T}^x_{-\lambda_2} \mathbf{T}^z_{-(\pi-\zeta)} \mathbf{T}^x_{-\lambda_1} \mathbf{T}^z_\xi \mathbf{T}^x_\phi \mathbf{T}^z_{(\pi-\alpha)} \mathbf{T}^x_\psi \mathbf{T}^z_\eta , \tag{9}$$

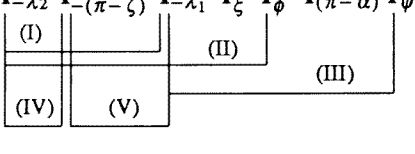
then the operation to derive eq. (4) can be viewed as a "contraction" between matrices $\mathbf{T}^x_{-\lambda_2}$ and $\mathbf{T}^x_{-\lambda_1}$, which in turn can be graphically represented by drawing a line connecting the two matrices:

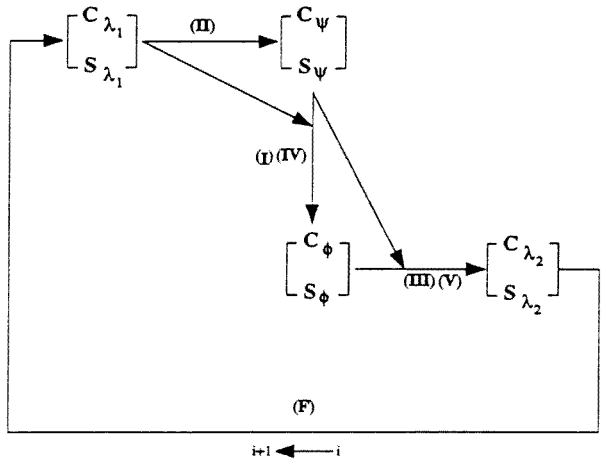
$$I = \underbrace{\mathbf{T}^x_{-\lambda_2} \mathbf{T}^z_{-(\pi-\zeta)}}\, \mathbf{T}^x_{-\lambda_1} \mathbf{T}^z_\xi \mathbf{T}^x_\phi \mathbf{T}^z_{(\pi-\alpha)} \mathbf{T}^x_\psi \mathbf{T}^z_\eta . \tag{10}$$

With this technique, one can easily see which two angular parameters are eliminated by a particular contraction, and choose the desired functional relations among the four variable parameters (or the corresponding four sines and four cosines).

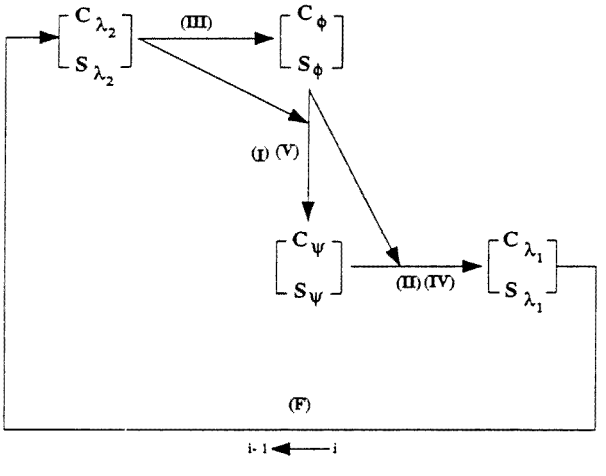
For example, the Purisima–Scheraga algorithm uses the following contractions:

$$I = \mathbf{T}^x_{-\lambda_2} \mathbf{T}^z_{-(\pi-\zeta)} \mathbf{T}^x_{-\lambda_1} \mathbf{T}^z_\xi \mathbf{T}^x_\phi \mathbf{T}^z_{(\pi-\alpha)} \mathbf{T}^x_\psi \mathbf{T}^z_\eta . \tag{11}$$





(a)



(b)

Fig. 3. The linkage map for the forward (a) and the backward (b) generations of the complete polypeptide backbone chain in the Purisima–Scheraga algorithm, where (I), (II), (III), (IV) and (V) are defined in eq. (11), (*F*) is eq. (2), and *S* = sine and *C* = cosine.

Here, the contractions (I), (II), (III), (IV) and (V) correspond to eqs. (9), (2), (8), (1) and (10) in [1]. The "linkage map", i.e. the flow-chart of the iterative procedures of generating successive backbone dihedral angles is shown in fig. 3.

4. An alternative algorithm: the λ-algorithm

One particular contraction, the one between $\mathbf{T}^{x}_{\phi}$ and $\mathbf{T}^{x}_{\psi}$, produces an equation:

$$(A): S_\xi S_\eta (C_\zeta C_{\lambda_1} C_{\lambda_2} + S_{\lambda_1} S_{\lambda_2}) + C_\eta S_\xi S_\zeta C_{\lambda_1} + C_\xi S_\eta S_\zeta C_{\lambda_2} + C_\alpha - C_\xi C_\eta C_\zeta = 0 \tag{12}$$

(where $S_\theta \equiv \sin\theta$ and $C_\theta \equiv \cos\theta$) that attracts special attention since it only involves two variables, λ_1 and λ_2. It is clear from eq. (2) that λ_1 and λ_2 are the only variables bridging across neighboring residues, while ϕ and ψ only relate to each other and to λ_1 and λ_2 within the same residue via eq. (1). It is the λ's that link the global structure of the polypeptide chain together in this geometric set-up.

This makes it possible to construct an algorithm, involving *only* variables λ_1 and λ_2, which iterates from one residue to the next and generates a "λ-chain" along

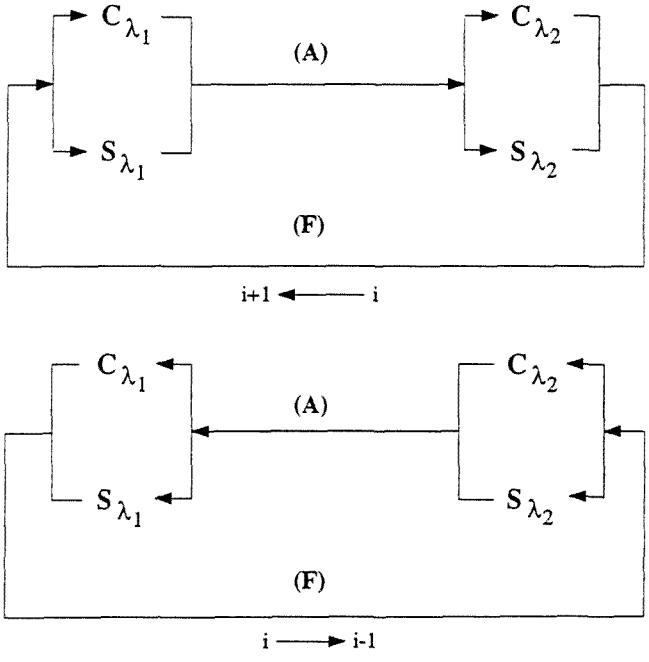


Fig. 4. The linkage maps for the forward (the upper diagram) and the backward (the lower diagram) generations of the λ-chain, where (A) is eq. (12) and (F) is eq. (2); $C_{\lambda_i} = \cos\lambda_i$, and $S_{\lambda_i} = \sin\lambda_i$, where $i = 1, 2$.

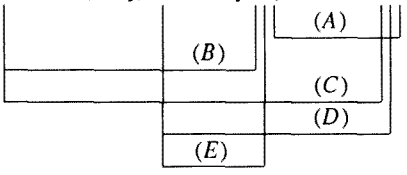
the given virtual-bond chain. The "linkage map" of such an algorithm is drawn in fig. 4. Note that since only eq. (12) relates λ_1 to λ_2, the orthogonality constraint

$$\sin^2\lambda_2 + \cos^2\lambda_2 = 1 \tag{13}$$

is also needed to determine $\sin\lambda_2$ and $\cos\lambda_2$ from $\sin\lambda_1$ and $\cos\lambda_1$ (or *vice versa*) of the same residue in the forward (or backward) generation. By substituting eq. (13) into eq. (12), $\sin\lambda_2$ can be eliminated and an equation quadratic in $\cos\lambda_2$ is obtained. For a given set of constant parameters $(\alpha, \xi, \eta, \zeta)$ and λ_1, there may be two, one or no solutions for $\cos\lambda_2$. Similar "branching" or "extinguishing" of solutions also happens in the PS algorithm, but there the branching point is chosen at converting $\cos\phi$ (or $\cos\psi$) to $\sin\phi$ (or $\sin\psi$) in the backward (or forward) generation [1].

To build up a real polypeptide backbone chain specified by the set of parameters $\{\phi, \psi\}$ from the λ-chain constructed by the above algorithm (fig. 4), other equations are needed. The following set of contractions seem to provide the simplest set of equations:

$$I = \mathbf{T}^x_{-\lambda_2}\,\mathbf{T}^z_{-(\pi-\zeta)}\,\mathbf{T}^x_{-\lambda_1}\,\mathbf{T}^z_{\xi}\,\mathbf{T}^x_{\phi}\,\mathbf{T}^z_{(\pi-\alpha)}\,\mathbf{T}^x_{\psi}\,\mathbf{T}^z_{\eta}\,, \tag{14}$$



where (A) is eq. (12). Two of the equations from this set of contractions are shown as follows:

$$(B):\ S_\alpha S_\eta C_\psi + S_\zeta S_\xi C_{\lambda_1} + C_\eta C_\alpha - C_\xi C_\zeta = 0, \tag{15}$$

$$(C):\ S_\alpha S_\zeta S_\phi S_{\lambda_1} - (S_\xi C_\alpha - C_\xi C_\phi S_\alpha) S_\zeta C_{\lambda_1} + C_\zeta (C_\xi C_\alpha + S_\xi C_\phi S_\alpha) - C_\eta = 0. \tag{16}$$

The other two equations (D) and (E) can be obtained from the above two by the following substitutions:

$$\begin{aligned} &\phi \leftrightarrow y, \\ &\xi \leftrightarrow \eta, \\ &\lambda_1 \leftrightarrow \lambda_2. \end{aligned} \tag{17}$$

This set of substitutions follow from the transformational properties of eq. (1) under transposition and the inversion of the signs of all angular parameters in the transposed matrix equation.

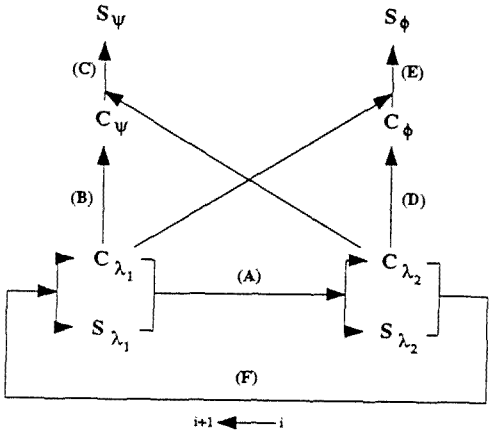


Fig. 5. The linkage map for the forward generation of the complete polypeptide backbone chain in the λ-algorithm, where (A), (B), (C), (D) and (E) are defined in eq. (14), (F) is eq. (2), and S = sine and C = cosine.

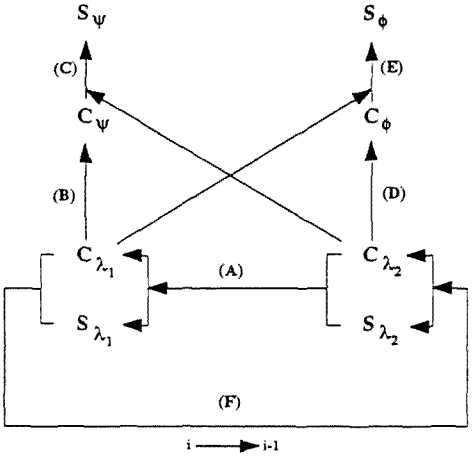


Fig. 6. The linkage map for the backward generation of the complete polypeptide backbone chain in the λ-algorithm, where (A), (B), (C), (D) and (E) are defined in eq. (14), (F) is eq. (2), and S = sine and C = cosine.

The linkage maps of the complete algorithm for converting a virtual-bond chain to a polypeptide backbone chain based on the choice of the equations in (14) are shown in figs. 5 and 6 for the forward and the backward generations, respectively. Note that the sines and cosines of ϕ and ψ of a residue are uniquely determined by the four linear equations (eqs. (15) and (16) and their two conjugate equations) from the λ_1 and λ_2 of the same residue. Since this algorithm is characterized by its

capability to generate the λ-chain $\{(\lambda_1), (\lambda_2)\}$ independently, and then to generate the real polypeptide backbone chain $\{\phi, \psi\}$ from the λ-chain separately, it is called the "λ-algorithm".

5. Discussion and summary

Based on the systematic analysis of the geometry for generating the full backbone of a polypeptide chain from a virtual chain of C^α coordinates, a new algorithm different from that of Purisima and Scheraga has been obtained. Actually, there exist many other different choices of the sets of equations relating the same set of geometric parameters in fig. 1, which may form the basis for different algorithms. However, it can be seen from the linkage-map construction that the λ-algorithm proposed here is the one involving the fewest steps in each iteration loop used to generate the necessary information for the next residue, i.e. the λ_2 in the forward generation and the λ_1 in the backward generation.

There are a number of similarities and differences between the λ-algorithm and the PS algorithm. Both algorithms are based on the same set-up of geometric parameters defined in figs. 1 and 2. The general analysis and the λ-algorithm proposed here are not aimed at problems beyond those which the Purisima–Scheraga algorithm* can deal with.

Both PS and the λ-algorithm use the same number of equations (five) derived from the fundamental geometric relations (1) and (2) and the same number (one) of orthogonality constraints. Therefore, the computation efficiency in generating all variables (ϕ, ψ, λ_1 and λ_2) within an iteration loop for one residue is about the same for the two algorithms. The differences between them become more apparent only at the level beyond one iteration loop.

In constructing the full polypeptide backbone by the virtual-bond chain method, there are two intrinsically different kinds of constraints working together to eliminate most of the solution branches. One is the "*local*" geometric and energetic constraint within the steric contacting range of one or a few neighboring residues, usually specified by some semi-empirical rules, e.g. the Ramachandran criterion; the other is the "*global*" geometric constraint, provided by the virtual-bond structure (through the set of virtual-bond chain parameters $\{\zeta, \gamma\}$), beyond the steric contacting range. The effects of the two types of constraints may be investigated separately by first neglecting the local constraint and trying to find all the branching paths of the solutions consistent with the given global structure of the virtual-bond chain, and then studying how these branching paths of the solutions can survive against the local constraints. The first part of the task, i.e. to search all the branching paths consistent with the global constraint, can be facilitated by the λ-algorithm since it

*For example, in practice, one may only be able to generate fragments of backbone from a protein virtual chain using this type of algorithm, because the regions between the fragments have conformations too different from those of the standard geometry to allow such algorithms to pass through.

involves only parameters $(\lambda_1)_i$ and $(\lambda_2)_i$. After all the "λ-chains" consistent with the given virtual-bond chain are found, the real polypeptide backbone chain (i.e. the Cartesian coordinates $\{x, y, z\}$) can be generated, each from a "λ-chain", and their consistency with the local constraints can then be checked.

The λ-algorithm may also improve the accuracy of the computation due to the fact that the calculations of ϕ_i and ψ_i are outside the iteration loop which consists of $(\lambda_1)_i$ and $(\lambda_2)_i$ only; therefore, the errors produced in computing ϕ_i and ψ_i of one residue will not propagate into the next iteration step for the computation of the neighboring (i.e. the $(i-1)$th or the $(i+1)$th) residue.

Acknowledgements

This work was partially supported by a grant for NASA (NSG-7305) and pursuant, in part, to a contract with the National Foundation for Cancer Research. N.L. is also supported in part by a grant (CA46838) from the National Cancer Institute. We are grateful to Dr. T.J. Zielinski for the informative discussions with her and for her critical reviews of the manuscript. We would also like to thank Dr. J. McDonald for useful discussions.

References

[1] E.O. Purisima and H.A. Scheraga, Biopolymers 23(1984)1207–1224.
[2] G.N. Ramachandran, C. Ramakrishnan and V. Sasisekharan, J. Mol. Biol. 7(1963)95–99.
[3] C. Ramakrishnan and G.N. Ramachandran, Biophys. J. 5(1965)909–933.
[4] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979),
[5] D. Poland and H.A. Scherga, *Theory of Helix-coil Transitions in Biopolymers* (Academic Press, New York, 1970).
[6] P.G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University, Ithaca, 1979).
[7] P.J. Flory, *Statistical Mechanics of Chain Molecules* (Interscience, New York, 1969).
[8] S. Srinivasan, M. Shibata, M. Roychoudhury and R. Rein, Int. J. Quant. Chem. QBS 14(1987) 281–288.
[9] J.J. McDonald and R. Rein, Int. J. Quant. Chem. QBS 16(1989)57–72.
[10] IUPAC-IUB Commission on Biochemical Nomenclature, Biochemistry 9(1970)3471–3479.
[11] F.A. Momany, R.F. McGuire, A.W. Burgess and H.A. Scheraga, J. Phys. Chem. 79(1975) 2361–2381.
[12] K. Nishikawa, F.A. Momany and H.A. Scheraga, Macromolecules 7(1974)797–806.