A Statistical Approach to the Calculation of Conformation of Proteins. 1. Theory

ARTICLE *in* MACROMOLECULES · JANUARY 1977

Impact Factor: 5.8 · DOI: 10.1021/ma60055a003 · Source: PubMed

1 AUTHOR:

Gordon Crippen
University of Michigan
150 PUBLICATIONS **5,810** CITATIONS

A Statistical Approach to the Calculation of Conformation of Proteins. 1. Theory

Gordon M. Crippen

Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143. *Received June 1, 1976*

A quantitative theory is presented for calculating the folding of a polymer, especially as applied to the renaturation of globular proteins. The theory applies to both thermodynamically and kinetically determined processes, so that not only can the equilibrium result be calculated but also the time course of folding. Since it is not automatically assumed that the folding will result in a single "native" conformation, it is possible to conclude that the polymer will exist as a random coil under appropriate conditions. A Monte Carlo method for applications is discussed in detail, and a very simple example is presented.

The question to be treated is how to calculate the equilibrium conformational state of a protein and/or the time course of the folding process. Of special interest is the formation of tertiary structure. The traditional view has been that the (unique) native state is the single conformation of globally minimal free energy (see, for example, ref 1 and 2). Attempting to globally optimize the energy of molecules larger than a dipeptide unfortunately leads to demonstrably insurmountable difficulties due to the numerous local minima.[3] Moreover, the customary description of conformations in terms of dihedral angles about rotatable bonds has made the study of tertiary conformation particularly difficult, since the long-range interresidue distances (and hence the energies) are very complicated and sometimes rapidly varying functions of dihedral angles. This has further aggravated the multiple minimum problem. In fact, searching for "the unique native structure" is justifiable only when the chosen external conditions are known to favor a single conformation, and otherwise it is reasonable to admit mixtures of partially folded states as outcomes of conformational calculations. Recently it has been suggested that the native state can be the most kinetically accessible product of the folding process,[4] or at least that kinetically favored side products are important.[5] In light of the above, a new approach is called for, one in which the objectives are redefined, conformations are differently described, and a statistical mechanical approach is used rather than a thermodynamic one.

Theory

Consider a protein solution sufficiently dilute that intermolecular interactions may be neglected. An individual molecule is assumed to take on a large but finite number of discrete conformations, according to the rotational isomeric model.[6] The molecule is self avoiding and not necessarily at Flory theta conditions.[6] There are a number of sites along the chain where energetically important interactions can occur, e.g., carbonyl and amide groups for forming hydrogen bonds, charged groups for salt bridges, alkyl groups for hydrophobic interactions, etc. These interactions are taken to involve only a pair of the sites each and then to be of importance only when the two sites are spatially close. Certainly hydrogen bonding, van der Waals forces, and solvation effects are short-distance phenomena, and it can even be argued that the presence of counterions makes salt bridges short distance also. Let the interactions corresponding to the various site parts be denoted by a single letter A, B, C, . . . , Z. Now any conformation may be described in terms of which interaction contacts it has. Clearly this description is closely related to the energy of the conformation, as long as all energetically important interactions have been included in list A, B, . . . Z, because one need merely add up the separate energies of the interactions to find the total interresidue energy. The individual conformations are not as important, however, as the classes of all conformations having a particular set of contacts but not any of the others. This is reasonable in light of the great conformational variability of some regions of the polypeptide chain in high resolution x ray crystal studies (see for instance ref 7 and 8). Clearly the interactions involving the variable parts of the chain are unimportant toward the description of an otherwise very well defined conformation. Let {A, C}, for example, denote the set of all conformations having contacts A and C but not B, D, E, . . . , Z. Let {∅} be the set of conformations having none of the chosen contacts. Thus the folding process may be expressed as a network of unimolecular reactions between these n classes of conformations. For example, one of the reactions might be {A} ⇌ {A, B}, whereby molecules already having contact A but not B close the gap to form B. Let $k_{AB,A}$ be the forward rate constant, $k_{A,AB}$ be the reverse rate constant, and $K_{AB,A} = (K_{A,AB})^{-1} = k_{AB,A}/k_{A,AB}$ be the equilibrium constant. Let c_A be the fraction of all molecules in the system which are in state {A}. Then the kinetic system may be formulated as the following set of n homogeneous, linear, first-order, differential equations in n variables:

$$\frac{dc_i}{dt} = k_{i,1}c_1 + \ldots + k_{i,i-1}c_{i-1} - \left[\sum_{\substack{j=1 \\ j\neq i}}^{n} k_{j,i}\right] c_i + k_{i,i+1}c_{i+1} + \ldots + k_{i,n}c_n \quad (1)$$

$$\text{for } i = 1, \ldots, n$$

where n is the number of classes, i.e., the number of molecular species being considered. This may be expressed more conveniently in matrix form:

$$\mathbf{c}'(t) = \Gamma\mathbf{c}(t) \quad (2)$$

where $\mathbf{c} = (c_1, \ldots, c_n)$ is the vector of mole fractions of the various molecular species, and the matrix Γ consists of elements

$$\gamma_{i,j} = \begin{cases} k_{i,j}, & i \neq j \\ -\sum_{\substack{l=1 \\ l\neq i}}^{n} k_{l,i}, & i = j \end{cases} \quad (3)$$

Equation 2 is then identical with the set of eq 1. In the case that all n eigenvalues of Γ are real and distinct, the solution $\mathbf{c}^*(t)$ to eq 2 is a linear combination of n terms of the form $\mathbf{v}_i e^{\lambda_i t}$, where $\mathbf{v}_i$ is the ith eigenvector of Γ and λ_i is the corre-

sponding eigenvalue. The linear combination is chosen so that $\mathbf{c}^*(0)$ equals the given initial mixture of conformations. If eigenvalues are degenerate or complex, the solution is somewhat more complicated, but it exists and is physically meaningful.[9] Numerical methods of solving eq 2 will be discussed in the Computational Considerations section. Alternatively, if one is interested only in the equilibrium state, then one need only have the equilibrium constant of each class with respect to one common class, say $\{\varnothing\}$. Then the $K_{A,\varnothing}$, etc., are the relative equilibrium mole fractions with respect to $c_\varnothing$, and they must only be normalized so that their sum is unity.

It remains to be explained how the rate constants and equilibrium constants are derived. For the reaction $\{A\} \rightleftarrows \{A, B\}$, calculate the equilibrium constant by

$$K_{AB,A} = \sum_{\{A,B\}} e^{-\epsilon/kT} \Big/ \sum_{\{A\}} e^{-\epsilon/kT} \tag{4}$$

Instead of the usual statistical mechanical treatment of integrating over phase space, we have summed over all conformations in the respective classes, since we have discretized the configuration space in using the rotational isomeric model of a polymer. Each ϵ is the internal energy corresponding to one conformation plus the free energy of the interaction with the solvent. We have in other words already integrated over all the degrees of freedom of the solvent. The rate constant $k_{AB,A}$ for the formation of contact B given contact A is taken to be independent of the relative energies of $\{A\}$ and $\{A, B\}$, because the energies of interaction were assumed to be significant only over short distances. Second, observe that $k_{AB,A}$ corresponds to the likelihood that interaction B will be added given the molecule is in state $\{A\}$. Remembering the customary definition of the conditional probability $p(\beta|\alpha)$ that event β will occur given the occurrence of event α

$$p(\beta|\alpha) = p(\beta \cap \alpha)/p(\alpha) \tag{5}$$

we are led to the analogous equation

$$k_{AB,A} = g_{AB}/(g_A + g_{AB}) \tag{6}$$

Equation 5 refers to the probabilities of events which are not necessarily mutually exclusive, while in eq 6 the states $\{A\}$ and $\{A, B\}$ are taken to be exclusive. The g's are the total number of conformations in the various classes, which is the discretized equivalent of the volume of conformation space occupied by a class. Physically speaking, we are claiming that the forward rate constant is governed predominantly by the relative librational entropies of the two states, and it is independent of the energy of the interaction being formed. Formation of a contact is taken to be a diffusion-like process where two parts of a chain come together as the result of a random walk, realizing any energetic advantage only when very close. Equation 6 has not been rigorously derived from a physical description of the diffusion process but rather by the analogy between rate constants and conditional transition probabilities. If experimental evidence required it, the definition of $k_{AB,A}$ could be altered by some multiplicative factor without affecting anything that follows. Let the rate constant $k_{A,AB}$ for breaking the contact B be defined by

$$K_{AB,A} = k_{AB,A}/k_{A,AB} \tag{7}$$

Thus the rate of breaking a contact does depend on the energy since energy enters into eq 4. Note that $k_{AB,A}$ and $k_{A,AB}$ are defined only up to a multiplicative constant, namely the Arrhenius factor containing the activation energy and a factor to determine the absolute time scale. Without experimental data (which is difficult to obtain for unimolecular reactions and for any sort of activation energy studies) one must speak only of *relative* rate constants, and then only when comparing similar interactions where the activation energies may be taken to be equal.

Example

Before discussing how the above theory may be applied in general, let us consider a very simple illustrative example. Let us simulate the conformations of a polypeptide of ten residues by a nine-step, self-avoiding walk on a two-dimensional square lattice. Suppose there are only two interactions of energetic importance: A between residues 1 and 10, and B between residues 2 and 5. Let them be equally important energetically and effective only when the two residues are on adjacent lattice points. For simplicity, suppose the formation of contacts to be irreversible, as would be the case at very low temperature, for instance. The classes of conformations to be considered are $\{\varnothing\}$, $\{A\}$, $\{B\}$, and $\{A, B\}$; take as possible reactions $\{\varnothing\} \rightarrow \{A\}$, $\{\varnothing\} \rightarrow \{B\}$, $\{A\} \rightarrow \{A, B\}$, and $\{B\} \rightarrow \{A, B\}$. Figure 1 shows one member of $\{A, B\}$. The first row of Table I gives the number of different conformations (up to translation and rotation in the plane, of course) for the various mutually exclusive classes, as determined by exhaustive enumeration. Applying eq 3 and 6, the Γ matrix results as shown. The upper triangle of Γ consists of zeros in this case because the reactions were assumed to be irreversible. The other elements were obtained straightforwardly, e.g., $\gamma_{21} = 0.0321 = 116/(116 + 3497)$. Such a simple problem can be solved exactly by first determining the eigenvalues as roots of the characteristic polynomial, $\det(\boldsymbol{\gamma} - \mathbf{I}\lambda) = 0$, where $\mathbf{I}$ is the identity matrix. In the present case of a lower triangular matrix, the eigenvalues λ_i are merely the diagonal elements of Γ, and they are listed in the far right column of the table. The corresponding unnormalized eigenvectors $\mathbf{v}_i$, $i = 1, \ldots, 4$, are given as row vectors beside them. The eigenvectors were found by solving $(\boldsymbol{\Gamma} - \mathbf{I}\lambda_i)\mathbf{v}_i = 0$. Let $\mathbf{c}^*(0) = (1, 0, 0, 0)$. This choice of initial conditions, along with our assumptions about the relative rate constants, leads us to simulate the irreversible folding when, say, a very good solvent is suddenly removed, so that initially all molecules were in relatively extended conformations. Then as explained in the previous section, we find the solution vector to be

$$\begin{aligned}
c_\varnothing^*(t) &= e^{-0.1416t} \\
c_A^*(t) &= 1.0772e^{-0.1416t} - 1.0772e^{-0.1714t} \\
c_B^*(t) &= -1.2345e^{-0.1416t} + 1.2345e^{-0.0529t} \\
c_{AB}^*(t) &= -0.8427e^{-0.1416t} + 1.0772e^{-0.1714t} \\
&\quad - 1.2345e^{-0.0529t} + 1
\end{aligned} \tag{8}$$

As shown in Figure 2, $\{B\}$ is present at all times in higher concentrations than $\{A\}$, which is what one would expect since residues 2 and 5 are so close in sequence and hence relatively likely to come in contact. Between times $t = 7$ and 14, $\{B\}$ predominates for kinetic reasons, but from then on $\{A, B\}$ becomes the dominant species, since it is thermodynamically the most favorable. The folding mechanism is essentially $\{\varnothing\} \rightarrow \{B\} \rightarrow \{A, B\}$. At equilibrium $c_{AB}^*(\infty) = 1$ and $c_\varnothing^*(\infty) = c_A^*(\infty) = c_B^*(\infty) = 0$.

Table I
Computational Tableau for Irreversible Folding of Lattice 10-mer

	{∅}	{A}	{B}	{A, B}		
g	3497	116	430	24		
Γ	−0.1416	0	0	0		
	0.0321	−0.1714	0	0		
	0.1095	0	−0.0529	0		
	0	0.1714	0.0529	0		
$\mathbf{v}_1$	1	1.0772	−1.2345	−0.8427	−0.1416	λ_1
$\mathbf{v}_2$	0	1	0	−1	−0.1714	λ_2
$\mathbf{v}_3$	0	0	1	−1	−0.0529	λ_3
$\mathbf{v}_4$	0	0	0	1	0	λ_4

Computational Considerations

If we are once able to calculate the matrix Γ, the equilibrium result may be similarly obtained via eq 4, and the kinetic results will be given by the vector function $\mathbf{c}^*(t)$, which is the solution of eq 2. Unfortunately, solving this equation involves finding in general all the eigenvalues and eigenvectors of the unsymmetric matrix Γ, which is not easy to do, even with approximate numerical methods. In many situations, the method cited by Hermans et al.[10] may be applied. For rough values of $\mathbf{c}^*(t)$, particularly when t is small, one may use the difference approximation to eq 2:

$$(\mathbf{I} + \Delta t\,\Gamma)^{\kappa}\mathbf{c}(0) \approx \mathbf{c}^*(\kappa\Delta t) \tag{9}$$

where $\mathbf{I}$ is the identity matrix, the iteration index $\kappa = 0, 1, 2, \ldots$, and $\Delta t > 0$ is chosen small enough that $1 + \Delta t\gamma_{i,i} \geq 0$ for all i, since otherwise there would be a chance that negative concentrations would result. The smaller Δt is, the more accurate the approximation.

Presuming conformational energies can be calculated well enough for use in eq 4, the remaining problem is to calculate the number of conformations g for each class for use in eq 6. The same problem is implicit in eq 4, because there each sum is to be taken over all the conformations in the given class. An exhaustive enumeration of all the possibilities, as in the example of the previous section, is out of the question for proteins, but remarkable success may be achieved by the following Monte Carlo method, as will be shown in the accompanying second paper in this series.[11] Basically one generates, at random, conformations belonging to a class and calculates for each conformation an estimate of the total number of such conformations. The average of the estimates is used as g in eq 6. In detail, let each of the n_r residues assume m conformations, according to the rotational isomeric model. For {A}, for example, generate conformations $j = 1, \ldots, n_A$ by choosing one of the m conformations at random for the first residue. With each subsequent residue i, try all m conformations, checking to see if each is allowed on steric grounds and that it is still possible to make the required long-range contacts while not making any of the others. (Contact checking is done whenever the first residue of a contact has already been generated, but the residue that must be close to it will be generated some k steps later. A necessary condition for making the contact is that k times the maximal distance coverable per residue is greater than the actual distance that must be traversed.) If all choices are disallowed, set the weight of the conformation, w_j, to zero and begin the next generation back at the first residue. Otherwise note $b_i \leq m$, the number of permissible conformations for residue i, choose one of them at random, and proceed to residue $i + 1$ until the entire jth random chain has been generated. Suppose that altogether

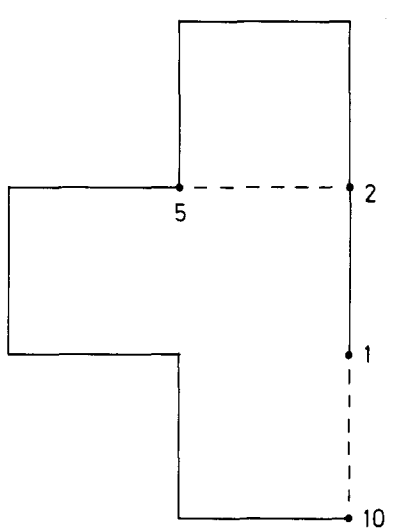

Figure 1. One example of a 10-mer on a square lattice in a class {A, B} = {1–10, 2–5} conformation.

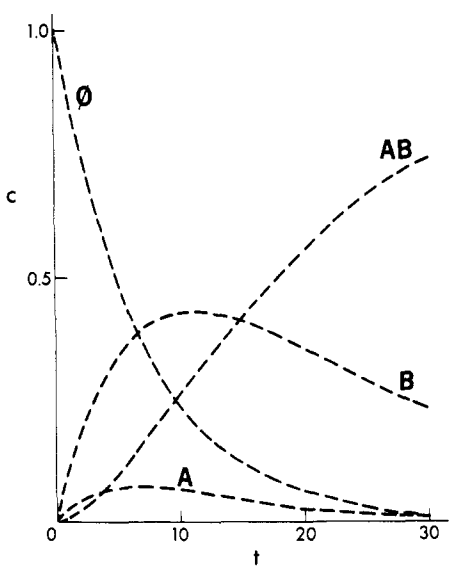

Figure 2. Analytical solution of the time course of the square lattice 10-mer folding when starting from relatively extended conformations. The concentration, i.e., population, c of each species is given as the fraction of the total number of molecules; the time t is in arbitrary units. ∅ = no contacts, A = 1–10 contact, B = 2–5 contact, and AB = both 1–10 and 2–5 contacts.

n_A such random chains belonging to class {A} have been generated. Then

$$w_j = m \prod_{i=2}^{n_r-1} b_i \tag{10}$$

is an unbiased estimator[12,13,14] of g_A, i.e.,

$$\lim_{n_A \to \infty} \sum_{j=1}^{n_A} w_j/n_A = g_A \tag{11}$$

The details of the above algorithm are crucial. *If* at the premature termination of a chain generation (because all m conformations of some intermediate residue are disallowed) one does *not* set $w_j = 0$ and include it in the average, then the estimation of g_A will be biased. If one does not begin the next chain generation back at the first residue after a premature or normal termination, the estimation of g_A will also be biased. See ref 14 for computational particulars. This generation method is not commonly used in Monte Carlo calculations on polymers because other investigators are usually interested in configurationally averaged quantities (e.g., mean square end-to-end distance) for very long chains, whereas here the concern is to find the number of conformations fulfilling certain requirements for chains of 100 or fewer residues. A test of eq 10 and 11 was made by generating 500 (including premature terminations) cyclic 10-mers on a square lattice, i.e., {A} in the example of the previous section using the above algorithm. The result was $g_A = 605 \pm 100$ vs. the exact value of $560 = 4(116 + 24)$ when rotations are included (and contact B is not excluded, as it is for {A} in Table I). The result is apparently not biased, although one would expect the estimation of g to be particularly bad when the restrictions on the random walk are strong. The stringency of the restrictions is certainly apparent in that there are $4^9 = 262\,144$ open, non-self-avoiding, 10-residue chains, but only 560 self-avoiding, cyclic 10-mers. The w's also remove the bias due to the sequential generation when estimating an averaged quantity. Thus we may use

$$K_{AB,A} = \lim_{n_{AB} \to \infty} \left[\frac{1}{n_{AB}} \sum_{j=1}^{n_{AB}} w_{AB,j} e^{-\epsilon_j/kT} \right] \Bigg/ \left[\lim_{n_A \to \infty} \frac{1}{n_A} \sum_{j=1}^{n_A} w_{A,j} e^{-\epsilon_j/kT} \right] \tag{12}$$

in place of eq 4. Physically speaking, $w_{A,j}$ is directly related to the librational entropy of conformation j in class {A}.

The variance in the estimation of the g's is unfortunately high, as can be seen from the test case in the previous paragraph. The following variance reduction method is extremely valuable, especially when the reactant and product classes are conformationally similar. Suppose we want to calculate $K_{AB,A}$ or $k_{AB,A}$, for example. For $j = 1, \ldots, n$ generate a conformation of {A} in the usual fasion, but also then generate a matching conformation of {A, B} by taking the same choices for each residue. When this is not possible for some residue, due to the additional requirement that interaction B be formed and the chain be still self avoiding, a suitable different conformation for the residue must be chosen at random. The result is n pairs of random chains, each member of the pair strongly resembling the other except for contact B. Then $w_{AB,j}/w_{A,j}$ is a lower variance estimator (i.e., lower standard deviation in a series of Monte Carlo trials) of g_{AB}/g_A than the equivalent expression for unpaired conformations or for independent samplings of {A} and {A, B}. Similarly, one can estimate $K_{AB,A}$ by

$$\sum_{j=1}^{n} w_{AB,j} e^{-\epsilon_j/kT} \Big/ \sum_{j=1}^{n} w_{A,j} e^{-\epsilon_j/kT}$$

or by averaging

$$w_{AB,j} e^{-\epsilon_j/kT} / w_{A,j} e^{-\epsilon_j/kT}.$$

Discussion

This approach to conformational calculations has a number of difficulties that must be dealt with in subsequent applications. (i) As with any conformational calculation, polypeptide geometry must be chosen, and an energy function must be devised for intra- and interresidue energies and solvation. (ii) In order to compare kinetic calculations with experiment, activation energies for various types of reactions and an absolute time scale must be taken from other experiments on unimolecular reactions. Not only are such experiments difficult, but although this theory envisages conformational change as a unimolecular process, the rate of disulfide bridge formation, for instance, has been shown to depend on the presence of many other factors in the solution.[15] (iii) In order to describe the renaturation of a protein from a rather extended, random-coil state, one might have to include a large number of interactions and a consequently much larger number of classes. Since a renaturing protein certainly does not have time to search all possible conformations, there must be certain restricted pathways of refolding, such as have been found for the reoxidation of reduced trypsin inhibitor.[16] Second, it has been observed in all the several protein crystal structures inspected that there are only an average of 19 interresidue close contacts per residue (C^α–C^α distance less than 10 Å).[17] We therefore propose that the refolding of most proteins can be described in terms of a few kinetically and/or thermodynamically favored intermediates and final products, and that these states can be adequately specified by a small number of contacts. (iv) This method requires the generation of many random conformations having certain given contacts. It is possible that there are no sterically allowed conformations that fulfill the requirements, or that the attrition in their generation (the premature termination of generation described earlier) will be great. For compact structures approaching the globular native state, this is an important problem, but it has been surmounted in some instances, as will be shown in the following paper.[11] (v) The greatest difficulty is the high variance, and hence low accuracy, in the estimation of the rate and equilibrium constants. Ordinarily the w's vary by only one or two orders of magnitude, so that pairing of large numbers of conformations makes estimation of the forward rate constants feasible. However, when the scatter in energies of the conformations is large compared to the temperature, the accuracy in determining the equilibrium constants and the reverse rate constants is very poor, even with pairing the conformations. The method appears to be applicable in practice only when either the variance in the energies is small or the two conformational classes being compared are very similar.

On the other hand, the approach has the advantage of en-

compassing both kinetic and thermodynamic determinism of admitting both microscopically diverse and uniform states as the time dependent or equilibrium description of the ensemble of protein molecules. Second, the designation of conformational states in terms of presence of interresidue contacts is not only inherently appropriate to the description of native conformations and directly related to the conformational energy, but it is also maximally economical. Not including some of the contacts would leave out energy contributions which are important by assumption. Extra contacts which occur in the conformations of a class but are not included in the designation of the class are geometric consequences of the explicit contacts. Third, the pitfalls of energy minimization are avoided altogether. Fourth, many experimental situations can be simulated. The experiment corresponding to the example calculation was sudden removal of guanidine hydrochloride at low temperature. Starting with the equilibrium mixture of states, the energy function can be altered to mimic a pH transition, or the temperature can be changed to simulate a thermal transition.

This method should not be considered a finished algorithm for solving all conformational calculation problems but rather a fresh conceptual framework to be built upon in future work. We have devoted much attention to its general applicability, lest it seem impractical. The next paper in this series will demonstrate its feasibility in a nontrivial problem.

References and Notes

(1) C. B. Anfinsen, *Science*, **181**, 223 (1973).
(2) H. A. Scheraga, *Chem. Rev.*, **71**, 195 (1971).
(3) G. M. Crippen, *J. Comput. Phys.*, **18**, 224 (1975).
(4) D. Wetlaufer, E. Kwok, W. L. Anderson, and E. R. Johnson, *Biochem. Biophys. Res. Commun.*, **56**, 380 (1974).
(5) C. Tanford, *Polym. Biol. Syst., Ciba Found. Symp.*, 124 (1972).
(6) P. J. Flory, "Statistical Mechanics of Chain Molecules", Interscience, New York, N.Y., 1969.
(7) F. S. Mathews, P. Argos, and M. Levine, *Cold Spring Harbor Symp. Quant. Biol.*, **36**, 387 (1971).
(8) C. C. F. Blake, D. F. Koenig, G. A. Mair, A. C. F. North, D. C. Philips, and V. R. Sarma, *Nature (London)*, **206**, 757 (1965).
(9) E. Kamke, "Differentialgleichungen—Lösungsmethoden und Lösungen", Vol. 1, 2nd ed, Akademische Verlagsgesellschaft, Leipzig, 1943.
(10) J. Hermans, Jr., D. Lohr, and D. Ferro, *Adv. Polym. Sci.*, **9**, 229 (1972).
(11) G. M. Crippen *Macromolecules*, following paper in this issue.
(12) M. N. Rosenbluth and A. W. Rosenbluth, *J. Chem. Phys.*, **23**, 356 (1955).
(13) J. M. Hammersley and D. C. Handscomb, "Monte Carlo Methods", Methuen, London, 1964.
(14) M. Janssens and E. DeVos, *J. Comput. Phys.*, **15**, 529 (1974).
(15) R. R. Hantgan, G. G. Hammes, and H. A. Scheraga, *Biochemistry*, **13**, 3421 (1974).
(16) T. E. Creighton, *J. Mol. Biol.*, **95**, 167 (1975).
(17) G. M. Crippen and I. D. Kuntz, unpublished results.

A Statistical Approach to the Calculation of Conformation of Proteins. 2. The Reoxidation of Reduced Trypsin Inhibitor[1]

Gordon M. Crippen

Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143.

Received June 1, 1976

ABSTRACT: A statistical method for the calculation of conformation is applied to the small protein, basic pancreatic trypsin inhibitor. The polypeptide geometry, energies, rotational isomers, and technique of conformation generation are discussed in detail. The calculations indicate that (i) under denaturing conditions, reduced trypsin inhibitor when oxidized should form initially 14 of the 15 possible single disulfide bridge intermediates in roughly equal proportions, and (ii) under renaturing conditions (pH 8.7, room temperature, aqueous solution) the single disulfide bridge intermediate with cys 30 and cys 51 connected is present in higher concentration than that with cys 30 and cys 55 linked. The two calculations are in agreement with experiment.

The previous paper in the series[2] discussed a novel statistical method of conformational calculation, which does not involve energy minimization. In this paper we describe the application of that method to the protein, basic pancreatic trypsin inhibitor. There are a number of reasons for choosing this compound for the first trials of the calculation technique. It is extremely small for a protein, having only 58 residues, yet its crystal structure is known and is exceptionally well defined.[3,4] The sequence is known,[5] and there are three disulfide bridges,[6] linking half-cystine residues 5–55, 14–38, and 30–51. There are no reduced cysteine residues in the native molecule. Recently Creighton has performed a series of studies[7–12] on the pathway of renaturation of reduced (essentially random coil[12]) trypsin inhibitor as it is allowed to reform its disulfide bridges in the presence of various oxidizing agents. Among many other things, he showed that the first intermediates to be formed had only a single disulfide bridge, and that they occurred in certain *approximate* relative concentrations: 50% 30–51 (a native bridge), 25% 5–30, 10% 30–55, 10% 5–51, and traces of other combinations.[9] This mixture of single disulfide intermediates is apparently an equilibrium phenomenon and can be reached either directly by reoxidizing the reduced protein under renaturing conditions as described or by reoxidation in 6 M guanidine hydrochloride, yielding all 15 possible