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

singly bridged compounds in roughly equal proportions, and then removing the denaturant, yielding the above mixture of predominantly only four intermediates.[13] In this paper, we test the capabilities of the statistical conformational calculations by showing that indeed the 30–51 species should be present in higher concentrations than 30–55 in an equilibrium mixture under renaturing conditions and that 14 of the 15 singly bridged intermediates should occur in roughly equal proportions under denaturing conditions. This is by no means a prediction by calculation, as the experimental results were known in advance. On the other hand, we will show it was no mere triumph of parameter adjusting.

Geometry and Energy

Before explaining the details of chain generation and averaging of conformational energies, we first state the simplified model of a polypeptide chain that was used and the corresponding energy function.

The protein molecule was reduced to a series of C^α atoms linked by virtual bonds and generated according to the method of ref 14. Although the N–C^α and C^α–C′ bonds are not explicitly included, the positioning of C^α_{i+1} depends on the ϕ and ψ values of residue i. The C^β atomsfor all residue types except glycine were placed in the appropriate orientation at 1.54 Å from the C^α. They were omitted entirely for glycine. Note that the location of C^β_{i+1} depends on ϕ_{i+1} as well as on ϕ_i and ψ_i. The location of all C^α and C^β atoms was then specified by giving ϕ and ψ for all n_r residues. The polypeptide chain was assumed to obey the rotational isomeric model,[14] where only certain ϕ, ψ combinations (corresponding to intraresidue energy local minima) are allowed. A set of combinations was selected for each of the 20 amino acid types from the work of Lewis et al.[15] They had minimized the energy of the *N*-acetyl-*N*′-methylamides of amino acids with respect to backbone *and* side chain conformation, so it was frequently necessary to lump together several conformations with similar ϕ and ψ but differing side chain conformation. It was necessary to choose at least one $\phi\psi$ pair in the α_R-helical region for each residue except proline and then alter its angles to a uniform $(\phi, \psi) = (-57°, -48°)$. See ref 16 for dihedral angle convention. Otherwise it was impossible for the computer program to generate several turns of α helix without some steric overlap, whereas in nature the angles could still be close to those of the rotational isomeric model, but slight deviations (depending on the amino acid sequence) would allow formation of helix. Not all choices of residue conformation were taken as equally probable, since some would correspond to narrow local minima of low statistical weight while others had broad minima of high weight. The calculated values in ref 15 of $(\det \mathbf{F})^{-1/2}$, where $\mathbf{F}$ is the matrix of second derivatives of the energy, were taken as the *relative* probability of occurrence of that conformation (see ref 17 for further explanation). Besides the one α-helical $\phi\psi$ choice, all conformations were taken that had a statistical weight (which is proportional[15] to $(\det \mathbf{F})^{-1/2}$) at 300 K of at least 0.3% of the total weights for that residue type. The result was: one choice for pro; three for arg; four for gln, glu, ile, phe, thr, trp, tyr, and val; five for ala, asp, cys, his, leu, lys, and met; six for asn and ser; and seven for gly. The choices for each residue except proline include α_R and extended conformations. Reasonably enough, asparagine, serine, and glycine, all frequently found in bends, are the most flexible. To each choice was assigned an intraresidue energy relative to zero energy for the energetically lowest choice for that residue type and a probability of occurrence normalized over all choices for that residue type. All these assignments were made before the trypsin inhibitor calculations were begun.

The energy of a given conformation was taken to be the sum of the individual intraresidue energies plus the pairwise interresidue energies. The intraresidue energy, in kcal/mol, of each allowed $\phi\psi$ choice for each residue type was taken from ref 15, as already mentioned. The interresidue energies consist of electrostatic, backbone–backbone, and hydrophobic interactions. The backbone–backbone part was a contribution of −0.1 kcal/mol (i.e., favorable) whenever two C^α atoms were closer than 6.5 Å. This term can be considered to simulate possible favorable dipole–dipole or hydrogen-bonding interactions between peptide groups compared to the alternative interaction between backbone and solvent. The much more important electrostatic energy term was computed according to the formula $332q_1q_2/(\epsilon r_{12})$, where q_1 and q_2 are the charges on fully ionized side chains (= ±1 electron charge units), r_{12} is the distance between the two C^β atoms in ångströms, ϵ is the dielectric constant, and 332 is a factor to convert the answer to kcal/mol. In the present calculation, we chose $\epsilon = 80$, the bulk dielectric constant of water, which is no doubt an overestimate for these small distances, but the effect of counterions was not otherwise included. The only charges included were those of the ionizable groups on side chains, the acidic ones having charge −1 above the pK_a of the group and zero charge when $pH < pK_a$, and similarly for the basic groups. Thus at pH 8.7, the conditions used experimentally,[9] a reduced cys residue is taken to be charged, but it is considered neutral when participating in a disulfide bond. The last part of the interresidue energy was the hydrophobic contribution. For simplicity, that was taken to be $(\Delta G_{tr1} + \Delta G_{tr2})/(0.1r_{12}^2)$ for each residue pair, where r_{12} is the distance in ångströms between the C^β atoms, and the ΔG_{tr}'s of residues 1 and 2 are the respective free energies of transfer (kcal/mole) from water to ethanol.[18] The ΔG_{tr} for the polar residues arg, asp, glu, and lys were arbitrarily set to +1 kcal/mol. The cys residues participating in a disulfide bridge are given the same ΔG_{tr} as phenylalanine, reflecting the low solubility of cystine in water. The arbitrary scaling constant 0.1 amounts to saying that if a residue were surrounded by 1.6 residues (regardless of type) at the minimal distance of 4 Å (excepting the two to which it is bonded) then the residue would be taken as completely immersed in the nonaqueous, alcohol-like medium. Note that for simplicity this term has neither a distance cutoff nor does it matter what kinds of residues surround the central residue. As the chain generation is self avoiding, there is also effectively a hard sphere infinite repulsive energy contribution whenever the C^α's of two residues come closer than 4 Å (the observed minimal distance in protein crystal structures). In summary, the geometry and energy of the protein have been enormously simplified, and many constants have been taken directly from other works. Although there are numerous opportunities for improving results by adjusting parameters and altering equations, we have experimented only with the dielectric constant. We shall also show that the energies are relatively unimportant for many of the results obtained.

Computational Methods

The general algorithm for random chain generation and subsequent Monte Carlo estimation of rate constants and equilibrium constants has already been explained in the previous paper of this series.[2] There are, howover, some refinements in the technique made necessary by the more sophisticated geometry of the present model.

Chain Generation. Using the earlier paper's notation,[2] the weight of the jth conformation, w_j, had been computed by $w_j = m\Pi_{i=2}^{n_r-1} b_i$, where n_r = number of residues, m = maximum number of allowed conformations of the first (and subsequent) residue, and b_i = number of allowed conformations of the ith residue ($b_i \leq m$). In this work we use instead

$$w_j = \prod_{i=2}^{n_r-1} \left(\sum_l p_l\right) \quad (1)$$

to reflect the variable probability, p, assigned to each choice of conformation. The sum extends over all allowed conformations of the residue i, and $\Sigma p \leq 1$ because the sum of *all* choices (whether allowed at some point in a generation or not), for a given residue type has been normalized to unity. Note that $w_j \leq 1$ is no longer itself the estimator of the total number of conformations but is proportional to it. Since we are always comparing two classes of conformations, the relative magnitude of w is all that matters. Chain generation is further complicated by the problem of high attrition, particularly when several interresidue contacts were required. This difficulty was solved by a "look ahead" algorithm reminiscent of chess playing programs:

(1) Place in fixed positions C_1^α, C_1^β, and C_2^α. Let $i = 3$.

(2) Choose, with equal likelihood, any of the allowed conformations of residue $i - 1$, thus fixing the locations of C^β_{i-1} and C^α_i.

(3) Check to see that (a) C^α_i is at least 4 Å away from any previously placed C^α; (b) C^α_i is so located that the second residue (yet to be generated) of a required contact could eventually be placed close to the first residue (already generated) of that same contact, allowing 3.0 Å/residue of chain progress toward that goal. The required maximum separation was 8 Å between either C^α's (as in β sheet backbone–backbone hydrogen bonding) or C^β's (as in a disulfide bridge).

(4) If either of the conditions (3a) or (3b) were not met, try the subsequent choice of conformation of residue $i - 1$ until one works. If none does, stop generation, set $w_j = 0$, and begin the next generation at step 1.

(5) If placement of C^α_i is successful, try all possible combinations of choices for residues i, $i + 1$, and $i + 2$ until the placements of C^α_{i+1}, C^α_{i+2}, and C^α_{i+3} are successful according to step 3. Ordinarily the first combination is approved. If no combination works, return to step 4.

(6) Having found a conformation of residue $i - 1$ that works and that leads to at least one successful generation of the next three residues, this conformation is irrevocably chosen. The remaining choices for residue $i - 1$ are briefly checked to see if placement of C^α_i is allowed, and the sum of the probabilities of all allowed choices is multiplied into the continued product of eq 1.

(7) Go to the next residue, increasing i by 1 and returning to step 2 until both $C^\alpha_{n_r}$ and $C^\beta_{n_r}$ have been placed, or the generation has been prematurely stopped in step 4.

One might object that this algorithm permits the generation of many very unlikely conformations by taking each residue conformational choice equally frequently in step 2, and thus overwhelming any configurational average with low weight (w_j) terms. The point, however, is to search out the extremely rare conformations that satisfy the numerous constraints of step 3 and compare them with similar, rare, constrained structures, rather than to average some quantity over the random-coil state.

Monte Carlo Estimation. The pertinent experiments on trypsin inhibitor are apparently equilibrium studies, so no rate constants have been calculated in this work. Equilibrium constants have been computed by basically the technique outlined in the previous paper,[2] where pairs of matching conformations were generated for the two classes to be compared. Averaging the terms $w_j \exp(-\epsilon_j/kT)$ was difficult because small variations in the energy produce large variations in these terms even though corresponding conformations of two classes were structurally similar. We have considered three different ways of averaging the terms so as to estimate $K_{A,B}$. (i) One method is the theoretically correct one, which we will call the "arithmetic average":

$$K_{A,B} \approx \sum_{j=1}^{n} w_{j,A} \exp(-\epsilon_{j,A}/RT) \Big/ \sum_{j=1}^{n} w_{j,B} \exp(-\epsilon_{j,B}/RT) \quad (2)$$

where n = number of pairs of conformations consisting of one from each of class {A} and {B}, and R is the perfect gas constant, used instead of Boltzmann's constant because the energies, ϵ, are in units of kcal/mol. Note that entries where $w_{j,A} = 0$ or $w_{j,B} = 0$ are included. (ii) We also have employed a geometric average, which is justifiable only in that it converges faster:

$$K_{A,B} \approx \left[\prod_{j=1}^{n'} w_{j,A} \exp(-\epsilon_{j,A}/RT)/w_{j,B} \exp(-\epsilon_{j,B}/RT) \right]^{1/n'} \quad (3)$$

In contrast to eq 2, the product is taken over the n' pairs where $w_{j,A} \neq 0$ and $w_{j,B} \neq 0$. The resulting estimation of $K_{A,B}$ tends to be closer to unity than it should be. (iii) Arithmetically averaging the ratios of the statistical weights, $w \exp(-\epsilon/RT)$, for corresponding pairs of conformations, gives rise to even greater variance than using eq 2, although it is not as uniformly dominated by terms having large w and small ϵ.

Results

As alluded to in the introductory section, calculations were performed to compare with two separate experimental findings: (1) at pH 8.7 and room temperature in aqueous salt solution, the renaturing intermediate of reduced trypsin inhibitor having the single disulfide bridge between cys residues 30 and 51 is present at approximately five times the concentration of the singly bridged species connecting 30 to 55; and (2) in 6 M guanidine hydrochloride, all 15 possible single disulfide combinations are present in roughly equal concentrations.

To simulate experiment 1, 100 attempts were made at generating trypsin inhibitor conformations with a required side chain–side chain contact between residues 30 and 51. Of these 100, all but 22 terminated prematurely. From the 22 successes, attempts were made to generate conformations with a required 30–55 side chain–side chain contact, of which 17 were successful. The result was 17 structurally similar pairs of conformations, one with 30–51 contact and the other with 30–55. Corresponding pairs of w's were markedly similar in spite of differing choices, so that although the w's within one class might differ by as much as a factor of 10^4, the average of their ratios for corresponding pairs was close to 1. Thus in comparing the two classes, the main consideration was the energy. Figure 1 shows an energy correlation diagram where the energies of the conformations of the 30–51 class are plotted along the horizontal axis against the energies of the corresponding conformations of the 30–55 class along the vertical axis. If the two classes had been energetically identical, all points would lie on the diagonal line. If there had been no resemblance in energies of corresponding pairs, the points would be scattered more uniformly over the figure. Observe that contact 30–51 conformations are significantly (i.e., a few kcal) favored over those with 30–55, that is, lie above the line in Figure 1, with only three relatively minor exceptions. In addition, the geometric averaged equilibrium constant = (concentration 30–55)/(concentration 30–51) = 0.08 (eq 3), and the arithmetic mean $K = 0.01$ (eq 2). For these averages, the temperature was taken to be 300 K, and the energies were all divided by 2 in order to improve the convergence of the averages. This did not affect the qualitative outcome.

To simulate the guanidine hydrochloride experiment, the same energy calculations were used as above, except that all charges and all free energies of transfer were set to zero. Then 20 conformations (out of 20 tries) of the completely reduced protein were generated as a standard to which we could compare each of the singly bridged classes and thence compare with each other. Accordingly, we generated corresponding sets of conformations for the 15 singly bridged classes, resulting in generation success ratios between 0/20 (for {5–14}) and 20/20

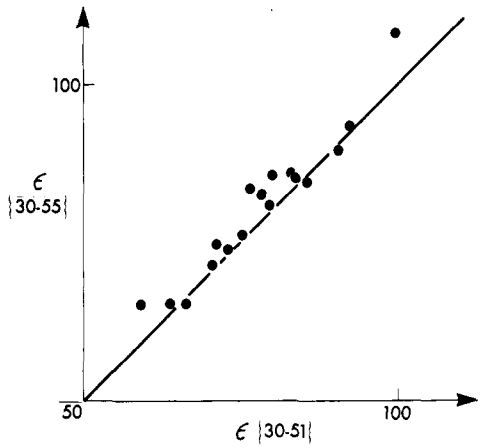

Figure 1. Energy, ϵ, in kcal/mol for 17 random conformations of trypsin inhibitor having only disulfide bridge 30–55 vs. the energies of the corresponding conformations having the bridge 30–51.

(for {51–55}). The one case of zero success, namely forming a 5–14 contact, is apparently possible by means of only *one* unique combination of conformations for residues 5 through 14, occurring only about once every 500 tries, as was discovered in another series of calculations. There are three proline residues (each having only one allowed conformation) in that short loop, which cause the difficulty in ring closure. The results, except for {5–14}, were that there was relatively little variation of w or ϵ within the individual classes and that the geometric averaged K's ranged between 0.3 and 2.8 relative to the fully reduced state. In other words, all 15 combinations should be present in comparable quantities, with the exception of perhaps {5–14}, which should have a particularly low concentration.

Conclusion

This statistical method of conformational calculation has been shown to be applicable to problems involving small proteins, particularly where the conformations involved are not very restricted. The poor convergence of the averages confine its use to situations where there is either little spread in the energies within a class or great conformational similarly between the classes. These conditions having been met in the two cases presented, the method was capable of results in agreement with, and accuracies comparable to, the semiquantitative[13] experiments. Work is now proceeding to adapt the procedure to a wider range of situations.

Acknowledgments. We kindly thank Drs. I. D. Kuntz, P. A. Kollman, and L. Peller for their helpful discussions concerning this work.

References and Notes

(1) This work was financially supported by the Academic Senate of the University of California.
(2) G. M. Crippen, *Macromolecules*, preceding paper in this issue.
(3) R. Huber, D. Kukla, A. Rühlmann, and W. Steigemann, *Cold Spring Harbor Symp. Quant. Biol.*, **36**, 141 (1971).
(4) J. Deisenhofer and W. Steigemann, "Proteinase Inhibitors", H. Fritz et al., Ed., Springer Verlag, Heidelberg, 1974, pp 484–496.
(5) J. Chauvet, G. Nouvel, and R. Archer, *Biochim. Biophys. Acta*, **115**, 130 (1966).
(6) F. A. Anderer and S. Hörnle, *J. Biol. Chem.*, **241**, 1568 (1966).
(7) T. E. Creighton, *J. Mol. Biol.*, **87**, 563 (1974).
(8) T. E. Creighton, *J. Mol. Biol.*, **87**, 579 (1974).
(9) T. E. Creighton, *J. Mol. Biol.*, **87**, 603 (1974).
(10) T. E. Creighton, *J. Mol. Biol.*, **95**, 167 (1975).
(11) T. E. Creighton, *J. Mol. Biol.*, **95**, 767 (1975).
(12) T. E. Creighton, *J. Mol. Biol.*, **95**, 777 (1975).
(13) T. E. Creighton, oral communication, 1975.
(14) P. J. Flory, "Statistical Mechanics of Chain Molecules", Wiley-Interscience, New York, N.Y., 1969.
(15) P. N. Lewis, F. A. Momany, and H. A. Scheraga, *Israel J. Chem.*, **11**, 121 (1973).
(16) J. C. Kendrew et al., *Biochemistry*, **9**, 3471 (1970).
(17) K. D. Gibson and H. A. Scheraga, *Physiol. Chem. Phys.*, **1**, 109 (1969).
(18) Y. Nozaki and C. Tanford, *J. Biol. Chem.*, **246**, 2211 (1971).

Conformation Dependent Transport Coefficients of Once-Broken Rods

Gerald Wilemski

Department of Chemistry, Dartmouth College, Hanover, New Hampshire 03755.
Received June 7, 1976

ABSTRACT: A method related to that of Kirkwood and Riseman is used to calculate the steady flow intrinsic viscosity as a function of the angle χ between the two arms of a once-broken rod. The translational diffusion coefficient and sedimentation coefficient are also obtained in terms of χ. The explicit χ dependence in the results allows for the likely possibility that the two arms of a real broken rod molecule will not be connected by a perfectly flexible joint and thus that the conformation of the molecule will be described by a nonrandom distribution of χ.

I. Introduction

Some years ago the once-broken rod was offered as a theoretical hydrodynamic model for incompletely helical polypeptides or other slightly flexible linear polymers, and the intrinsic viscosity[1] $[\eta]$ and translational diffusion coefficient D (or sedimentation coefficient s)[2,3] were calculated. At about the same time, independent experimental studies[3] of the solution viscosity of poly(γ-benzyl L-glutamate) (PBLG) containing a flexible trimethylene diamine unit gave results in apparently excellent agreement with the calculated intrinsic viscosity, as compared to that for an unbroken rodlike polypeptide.

In fact, the situation requires further study. On the theoretical side, Hassager[4] has pointed out an error in some of the old calculations[1] and has corrected the result for the free-draining steady flow intrinsic viscosity. Furthermore, because the samples studied were polydisperse and possibly contaminated with unbroken rods, the experimental data[3] are also less than definitive. However, the dielectric measurements