# Conformational Analysis of Biologically Active Polypeptides, with Application to Oncogenesis

MATTHEW R. PINCUS*

*Department of Pathology, New York University Medical Center, New York, New York 10016*

HAROLD A. SCHERAGA*

*Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853-1301*

*Received April 30, 1985 (Revised Manuscript Received September 23, 1985)*

Proteins and biologically active polypeptides perform many different functions, all of which depend on their three-dimensional structures. It is currently accepted that the structure adopted by a given polypeptide chain is determined by its amino acid sequence and its environment.[1] Changes in the sequences or environments of polypeptides or proteins can lead to radical changes in their structures and consequently in their functions. A dramatic example of this phenomenon involves the P21 proteins (proteins of molecular weight 21 000) coded for by oncogenes and thought to be involved in the regulation of the cell cycle in all eukaryotic cells.[2] It has recently been observed that a single amino acid substitution at position 12 of a P21 protein (or at a number of other positions along the polypeptide chain) causes the protein to become oncogenic, i.e., to induce malignant transformation of normal cells.[3-5] It is, therefore, desirable to gain an understanding of the factors that determine the structures and functions of polypeptides and proteins and of how they may be altered by changes in amino acid sequence.

We have developed methods, based on conformational energy calculations, to determine the preferred structures of long polypeptide chains and simple proteins of biological importance.[6-10] Many of the predicted structures have been confirmed experimentally by X-ray crystallography and/or by solution studies, mainly by NMR. Other predicted structures have been corroborated in biological and genetic experiments as discussed below. As a result, it is now possible to determine the interactions that stabilize biologically active structures of polypeptides and proteins and to analyze how amino acid substitutions affect these structures.

This methodology has also been extended to more complex systems. In a previous Account,[11] we showed how these computer-based methods can be used to predict the three-dimensional structures of enzyme–substrate complexes. In the case of lysozyme, for example, we deduced modes for the binding of polysaccharide substrates to the enzyme that were not previously known but were subsequently confirmed by experiments, including the use of a competitive binding assay between the substrate and a site-specific monoclonal antibody.[12]

In this Account, we discuss the results of conformational analysis of some illustrative constrained polypeptides (such as gramicidin S) and of unconstrained membrane polypeptides (such as melittin). We also present an analysis of the preferred conformations for the transforming region of the oncogenic P21 proteins[10] and demonstrate how single amino acid substitutions in a normal protein can cause major local changes in protein conformation. These latter results illustrate how theoretical conformational analysis can be used to gain insights into processes involved in oncogenesis.

## Methods Used in Conformational Energy Calculations of the Preferred Structures of Polypeptides and Membrane Proteins

The methods used to compute the preferred structures of polypeptides have been discussed extensively elsewhere.[6,7] In brief, the conformational energy of a polypeptide is computed by using the program ECEPP/2.[13,14] This energy is given by the expression

$$U = \sum_{i<j} \epsilon_{ij} \left[ \left( \frac{r_{ij}^{0}}{r_{ij}} \right)^{12} - 2 \left( \frac{r_{ij}^{0}}{r_{ij}} \right)^{M} \right] + \sum_{i<j} \frac{q_i q_j}{D r_{ij}} + \sum_{k} \frac{A_k}{2} (1 \pm \cos n\theta_k) \quad (1)$$

Matthew R. Pincus was born in Brooklyn, NY, in 1948. He received his M.D. and Ph.D. in Biochemistry from SUNY Downstate Medical Center, was a postdoctoral fellow with Professor Scheraga at Cornell, completed his residency in Pathology at Columbia Presbyterian Medical Center in New York City, was an Expert-Consultant in the Laboratory of Theoretical Biology at NIH, and is currently Professor of Pathology at New York University Medical Center.

Harold A. Scheraga was born in Brooklyn, NY, in 1921. He attended the City College of New York, where he received his B.S. degree, and went on to graduate work at Duke University, receiving the Ph.D. in 1946 and, in 1961, an Sc.D. (Hon.). Following postdoctoral work at Harvard Medical School, he joined the faculty at Cornell University where he is Todd Professor of Chemistry. His research interests are in the physical chemistry of proteins and other macromolecules, chemistry of blood clotting, and structure of water and dilute aqueous solutions.

(1) C. B. Anfinsen and H. A. Scheraga, *Adv. Prot. Chem.*, **29**, 205-300 (1975).
(2) E. Taparowsky, K. Shimizu, M. Goldfarb, and M. Wigler, *Cell*, **34**, 581-586 (1983).
(3) C. J. Tabin, S. M. Bradley, C. I. Bargmann, R. A. Weinberg, A. G. Papageorge, E. M. Scolnick, R. Dhar, D. R. Lowy, and E. H. Chang, *Nature (London)*, **300**, 143-149 (1982).
(4) E. P. Reddy, R. K. Reynolds, E. Santos, and M. Barbacid, *Nature (London)*, **300**, 149-152 (1982).
(5) D. J. Capon, E. Y. Chen, A. D. Levinson, P. H. Seeberg, and D. V. Goeddel, *Nature (London)*, **302**, 33-37 (1983).
(6) G. Némethy and H. A. Scheraga, *Q. Rev. Biophys.*, **10**, 239-352 (1977).
(7) H. A. Scheraga, *Carlsberg Res. Commun.*, **49**, 1-55 (1984).
(8) M. R. Pincus and R. D. Klausner, *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 3413-3417 (1982).
(9) M. R. Pincus, R. D. Klausner, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 5107-5110 (1982).
(10) M. R. Pincus, J. van Renswoude, J. B. Harford, E. H. Chang, R. P. Carty, and R. D. Klausner, *Proc. Natl. Acad. Sci. U.S.A.*, **80**, 5253-5257 (1983).
(11) M. R. Pincus and H. A. Scheraga, *Acc. Chem. Res.*, **14**, 299-306 (1981).
(12) S. J. Smith-Gill, J. A. Rupley, M. R. Pincus, R. P. Carty, and H. A. Scheraga, *Biochemistry*, **23**, 993-997 (1984).

0001-4842/85/0118-0372$01.50/0 

where $\epsilon_{ij}$ and $r_{ij}^0$ are the potential well depth and position of the minimum of the pair interaction energy (for nonbonded energy, $M = 6$; for hydrogen-bonding energy, $M = 10$), $q$ is the partial atomic charge, $D$ is the dielectric constant, $r_{ij}$ is the distance between the two interacting atoms, $A_k$ is the barrier height for rotation around the $k$th bond, $\theta_k$ is the dihedral angle, and $n$ is the $n$-fold degeneracy of the torsional potential.

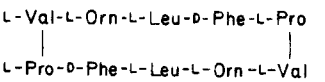
These potential functions have been used to compute the single-residue minima for all of the naturally occurring amino acids.[15-17] They have then been applied to compute the preferred conformations of oligopeptides (di- through hexapeptides)[6,7,18,19] and oligo- and polysaccharides.[20] For the biologically active molecules among these, the predicted structures have correlated quite well with experimentally determined structures (see, for example, ref 6, 7, 19, and 20). These computations of the structures of short (oligo-) peptides have been performed in a systematic way involving a directed search in which all possible combinations of the single-residue minima for the first two amino acids in a given sequence are generated and subjected to energy minimization.[21] Dipeptide conformations whose energies lie within a certain cutoff energy value above that of the global (lowest energy) minimum are retained and then combined with all of the single residue minima for the next residue, and the entire build-up process described above[22] is repeated. Variations of this procedure have also been used.[23] This approach is feasible for computing the energetically preferred conformations for oligopeptides but encounters difficulties with longer polypeptides because, even with the use of single-residue minima and energy cutoffs, the number of low-energy conformations for polypeptides (using this chain build-up method) increases greatly, making further computations technically difficult. Nevertheless, this method has been used successfully in several cases, e.g., gramicidin S[24] and the triple-stranded collagen model poly(Gly-Pro-Pro).[25] Both of these important biological molecules contain natural constraints, the first because it is a cyclic peptide with symmetry and the second because it is a polymer with perfectly repeating units.

It should be noted that, in the polypeptide calculations presented here, the effects of solvent have been omitted because the structures of these polypeptides very likely are determined by intramolecular interactions. Gramicidin S and poly(Gly-Pro-Pro) contain important intramolecular constraints and interactions that are critical in influencing folding. Also, the computations have been applied to membrane polypeptides and proteins, such as melittin, which are known to require nonpolar environments to fold properly; in these cases, water actually *denatures* the peptides. In other calculations (not reported here) the effect of water is taken into account either with a solvent-shell model or with statistical data from protein X-ray structures.[6,7]

As will be illustrated in the examples below, conformational analysis of the different polypeptides considered here yields at most several very similar structures whose conformational energies are much lower than those of possible "competing" structures. Thus, the effect of conformational entropy is not likely to influence the preferred conformations of these peptides. In the few cases in which two or more structures differ by small energies, it has been found that they are sufficiently similar in structure that their conformational entropies would not differ. In those cases in which entropy is expected to make a significant contribution, it is included.[6,7]

## Gramicidin S

This cyclic decapeptide antibiotic has the sequence:




Computation of the allowed conformations of this polypeptide was facilitated by the facts that the sequence is repetitive and the peptide is cyclic. The cyclic property imposes a constraint in that all structures generated for the molecule must satisfy a ring-closure condition. Use of the methods described above, combined with the ring-closure condition,[24] required the minimization of the energies of over 10 000 starting conformations. Of the resulting energy-minimized conformations, only one lowest energy or global-minimum-energy structure, shown in Figure 1A, was obtained; it was approximately 2 kcal/mol lower in energy than the next lowest energy structure.[24] Calculations of the expected $^1$H NMR coupling constants for side chains and backbone vicinal proton pairs was in excellent agreement with available NMR data for this molecule.[24] Subsequently, the structure of gramidicin S was determined by X-ray crystallography[26] (Figure 1B) and found[27] to be quite similar to the calculated one (compare parts A and B of Figure 1). Raman and infrared studies, interpreted with the aid of normal mode calculations,[28] also support this structure. As may be seen in both structures (Figure 1A,B), there is an extensive network of intramolecular NH···O=C hydrogen bonds (dotted lines) across the chain. The two Val-Orn-Leu sequences adopt an antiparallel $\beta$-pleated sheet conformation with respect to one another while the two Phe-Pro sequences form $\beta$-turns. In addition, in each structure, an important hydrogen bond forms between the $\delta$-$NH_2$ group of Orn (not one of the naturally occurring amino acids but one for which the sin-

(13) F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79**, 2361–2381 (1975).
(14) G. Némethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.*, **87**, 1883–1887 (1983).
(15) P. N. Lewis, F. A. Momany, and H. A. Scheraga, *Israel J. Chem.*, **11**, 121–152 (1973).
(16) S. S. Zimmerman, M. S. Pottle, G. Némethy, and H. A. Scheraga, *Macromolecules*, **10**, 1–9 (1977).
(17) M. Vásquez, G. Némethy, and H. A. Scheraga, *Macromolecules*, **16**, 1043–1049 (1983).
(18) S. S. Zimmerman and H. A. Scheraga, *Biopolymers*, **16**, 811–843 (1977).
(19) I. Simon, G. Némethy, and H. A. Scheraga, *Macromolecules*, **11**, 797–804 (1978).
(20) M. R. Pincus, A. W. Burgess, and H. A. Scheraga, *Biopolymers*, **15**, 2485–2521 (1976).
(21) M. J. D. Powell, *Comput. J.*, **7**, 155–162 (1964).
(22) H. A. Scheraga, *Biopolymers*, **22**, 1–14 (1983).
(23) M. Vásquez and H. A. Scheraga, *Biopolymers*, **24**, 1437–1447 (1985).
(24) M. Dygert, N. Gō, and H. A. Scheraga, *Macromolecules*, **8**, 750–761 (1975).
(25) M. H. Miller and H. A. Scheraga, *J. Polym. Sci., Polym. Symp.*, **54**, 171–200 (1976).
(26) S. E. Hull, R. Karlsson, P. Main, M. M. Woolfson, and E. J. Dodson, *Nature (London)*, **275**, 206–207 (1978).
(27) S. Rackovsky and H. A. Scheraga, *Proc. Natl. Acad. Sci., U.S.A.*, **77**, 6965–6967 (1980).
(28) V. M. Naik, S. Krimm, J. B. Denton, G. Némethy, and H. A. Scheraga, *Int. J. Peptide Protein Res.*, **24**, 613–626 (1984).

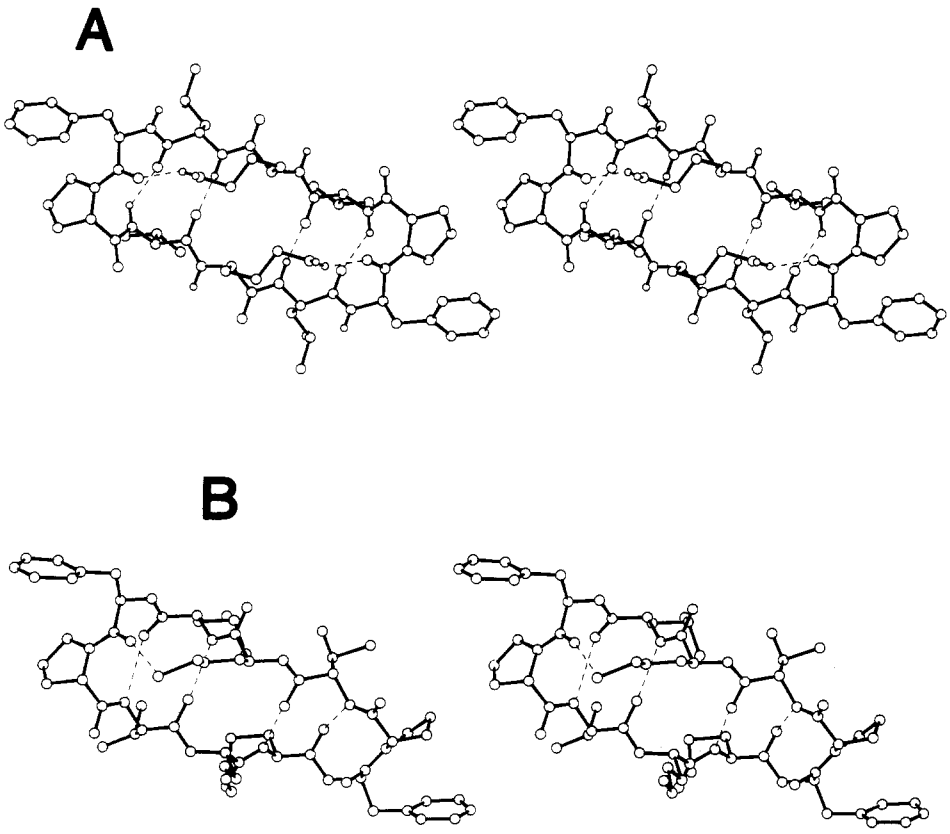



**Figure 1.** A. Stereoview of the computed lowest energy structure for gramicidin S,[24] showing the hydrogen-bonding network (dashed lines) that stabilizes this structure. B. Stereoview of the X-ray structure of gramicidin S.[26]

gle-residue minima have been determined) and the C=O of a preceding D-Phe residue. (The Orn-Phe hydrogen-bonding pairs are equivalent since the final structures of both possess $C_2$ symmetry.[29]) Thus, specific hydrogen bonds appear to stabilize the structure of this molecule. Possible substitution of L-amino acids for D-Phe and amino acids other than Lys for Orn in this peptide would be expected to destabilize this structure and possibly alter its function.

## Collagen

This fibrous protein has been shown to be composed of triple helices formed by three strands of single polypeptide chains whose sequences are of the form poly(Gly-X-Y).[30] The methods described above were used to compute the preferred conformations for poly(Gly-Pro-Pro).[25] First, all of the low-energy minima for Gly-Pro-Pro were determined (the conformational space of this unit being highly constrained by the few preferred conformations for Pro). The allowed conformations for this basic unit could then be combined to generate the preferred conformations for any desired chain length, assuming that all tripeptide units in the chain are equivalent. For a single chain of $(Gly\text{-}Pro\text{-}Pro)_4$, there were numerous low-energy conformations. However, when each of these was packed into triple-stranded helices satisfying various symmetry constraints (assuming that the three chains are equivalent), remarkably only one lowest energy structure (a helix) was found, which was 19 kcal/mol lower in energy than the structure of the next lowest energy.[25] After these calculations were completed, the X-ray crystal structure of $(Gly\text{-}Pro\text{-}Pro)_{10}$ was determined.[31] Comparison of the atomic coordinates of the computed global-minimum-energy structure with those of the X-ray structure revealed a root-mean-square deviation of only 0.3 Å between them; i.e., the structures were superimposable.[25] The packing of the three chains is such that they coil around one another, i.e., form "coiled coils". The chains are tightly packed. Hydrogen bonds form between the NH of the Gly residue of one chain and the C=O of the Pro residue of a neighboring chain. Similar calculations have been performed for poly(Gly-Pro-Hyp)[32] and poly(Gly-Pro-Ala),[33] both of which adopt the collagen-like coiled-coil conformation found for poly(Gly-Pro-Pro).[25] However, poly(Gly-Ala-Pro)[34] does not adopt this structure as its lowest energy conformation, indicating sensitivity of structure to specific sequence.[34]

The long-range interactions that exist between chains in collagen, from the tight packing of chains (involving many weak nonbonded interactions described by the first sum in eq 1, adding up to produce significant stabilization energy) and from specific hydrogen bonds, result in selection of only one conformation from many available ones for the single chain. Such a process is probably involved in the folding of globular proteins where two or more sequences of the same chain interact so that only one of many possible conformations for each sequence is selected.

(29) G. Némethy and H. A. Scheraga, *Biochem. Biophys. Res. Commun.*, 118, 643–647 (1984).
(30) W. Traub and K. A. Piez, *Adv. Protein Chem.*, **25**, 243–352 (1971).
(31) K. Okuyama, N. Tanaka, T. Ashida, and M. Kakudo, *Bull. Chem. Soc. Jpn.*, **49**, 1805–1810 (1976).
(32) M. H. Miller, G. Némethy, and H. A. Scheraga, *Macromolecules*, 13, 470–478 (1980).
(33) M. H. Miller, G. Némethy, and H. A. Scheraga, *Macromolecules*, 13, 910–913 (1980).
(34) G. Némethy, M. H. Miller, and H. A. Scheraga, *Macromolecules*, 13, 914–919 (1980).

Thus, application of the build-up procedure for constrained polypeptides, as described above, results in predictions of structures that are in agreement with experiment. The method can even be extended to incorporate multiple interchain interactions as in the case of collagen. Because the chain build-up procedure has been successful on oligopeptides and constrained polypeptides, it was desirable to extend the method to long, unconstrained polypeptides. The build-up procedure, however, when extended to nonregular sequences that have no constraints, generates very large numbers of low-energy conformations for short peptide segments making repeated application of the method very difficult.

## Structures of Long Polypeptides and Membrane Proteins

It has been possible to extend these methods to unconstrained long polypeptide sequences and to circumvent the problems mentioned above. When the many minima for oligopeptide segments are examined, it is noted that many of them have essentially identical backbone conformations but different rotational isomeric states for their side chains. Thus, many conformations are "over-specified" and only those with uniquely different backbone conformations need to be considered in the chain build-up procedure. Combination of the unique or nondegenerate minima for neighboring peptide segments in a polypeptide chain then becomes feasible and allows computation of the allowed structures for long polypeptide sequences. This method is called the method of combination of nondegenerate minima[8,9] and has been applied to a number of membrane polypeptides and proteins known to require membranes or nonpolar environments to fold correctly.

## Leader Sequences

Almost all secreted proteins contain sequences of 15–30 amino acid residues on their amino termini called leader sequences or signal peptides that function to cause translocation of the protein chain across membranes.[35] Once translocation has occurred, the leader sequences are rapidly cleaved from the protein. The amino acid sequences of many of these signal peptides are known and differ among various secreted proteins. The fact that many of these proteins are translocated across the same membrane suggest that, despite differences in sequence, the signal peptides possess common structural features that allow protein secretion. All leader sequences have long segments of contiguous nonpolar amino acid residues, and it is these sequences that may adopt the structure common to all of the signal peptides. The presence of these long nonpolar peptides suggests that they may act as chain-folding initiation sites (CFIS),[36] i.e. as sequences that adopt a limited number of conformations and cause neighboring sequences to adopt compatible conformations.

The preferred conformations of the leader sequence of murine pre-$\kappa$-light chain were explored by using the methods described above.[8] This signal peptide has the sequence Met-Glu-Thr-Asp-Thr-[(Leu)$_3$-Trp-Val]$_2$-Pro-Gly. (The Met functions as a chain initiating amino acid and was not included in the computations, although it is referred to here as residue no. 1.) The preferred conformations for the repeating nonpolar decapeptide were determined first. Of many possible starting conformations, this sequence adopted an $\alpha$-helical conformation as the global minimum. All low-energy conformations were $\alpha$-helical from Leu-6 to Val-15, the last two residues (Trp-Val) in this fragment adopting a number of different conformations.[8] When these conformations were added to those determined for Pro-Gly, a unique lowest energy structure was obtained that was $\alpha$-helical from Leu-6 to Trp-14 while the Val-Pro-Gly sequence terminated the helix in a distinct twist conformation. The initial polar tetrapeptide (Glu-2 to Thr-5) adopted a nonhelical, mainly $\beta$-conformation when its low-energy structures were combined with that for the Leu-6-Gly-17 sequence. Thus, the nonpolar segment of this leader sequence does adopt a unique conformation, an $\alpha$-helix. This result was further confirmed by placing the nonpolar decapeptide in a number of different $\beta$-pleated sheet conformations such that the two (Leu)$_3$ sequences were extended and antiparallel to one another while the central Trp-Val was generated in a number of different $\beta$-bend conformations. Energy minimization of all of these structures yielded no structure whose energy was less than 10 kcal/mol higher than that for the $\alpha$-helical structure.[8]

The results that most of the nonpolar sequence exists as an $\alpha$-helix and that this helix comprises about 50% of the molecule have been confirmed by a number of different experiments on leader sequences. In one set of genetic experiments, the helix-breaking residue $\beta$-hydroxyleucine was introduced in place of leucine into a growing leader sequence during protein synthesis.[37] Incorporation of this residue into the leader sequence greatly reduced the amount of protein translocation.[37] In other genetic experiments involving recombinant DNA in bacterial systems, deletion of several nonpolar residues from a leader sequence was accomplished.[38] The resulting leader sequences had nonpolar sequences with two Pro residues (Pro being a strong $\alpha$-helix breaker) within four amino acid residues of one another. These mutants were likewise found to be translocation defective.[38] Physical studies, particularly circular dichroism, on a number of different synthetic leader sequences including pre-pro-parathyroid hormone[39] in non-hydroxylic solvents all indicate helix contents in the range of 40–50% which is approximately the fraction of the residues of the signal peptides that comprise the nonpolar segment. The twist conformation that terminates the helix may be a general structural feature recognized by endopeptidases in the membrane that cleave the signal peptide.[8]

## Melittin

This membrane-active protein is present in bee venom and binds to cell membranes, causing trans-membrane channels to form with consequent extrusion of

(35) G. Blobel and B. Dobberstein, *J. Cell Biol.*, **67**, 835–851 (1975).
(36) G. T. Montelione, E. Arnold, Y. C. Meinwald, E. R. Stimson, J. B. Denton, S. G. Huang, J. Clardy, and H. A. Scheraga, *J. Am. Chem. Soc.*, **106**, 7946–7958 (1984).
(37) G. Hortin and I. Boime, *Proc. Natl. Acad. Sci. U.S.A.*, **77**, 1356–1360 (1980).
(38) S. D. Emr, M. N. Hall, and T. J. Silhavy, *J. Cell Biol.*, **86**, 701–711 (1980).
(39) M. Rosenblatt, N. V. Beaudette, and G. D. Fasman, *Proc. Natl. Acad. Sci. U.S.A.*, **77**, 3983–3987 (1980).

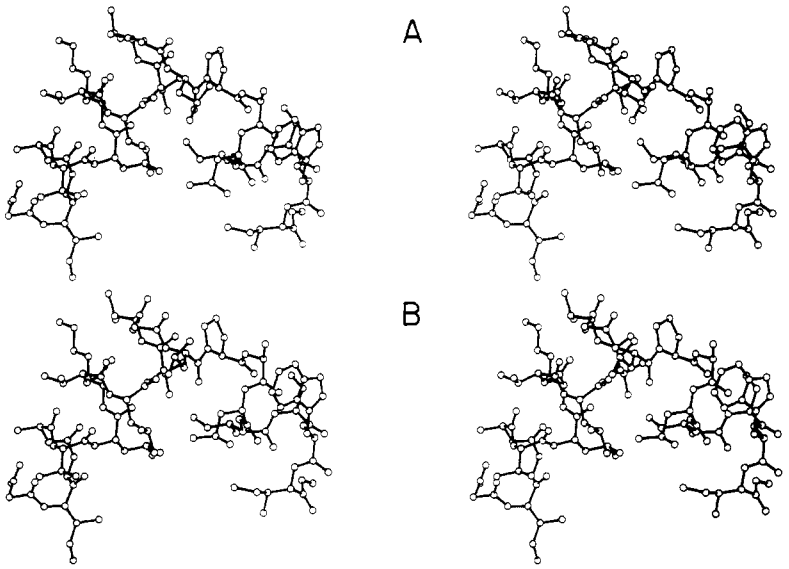



**Figure 2.** Stereoview of the two computed lowest energy structures of residues 1–20 of melittin (2A,B),[9] showing packing of nonpolar residues in the interior of the molecule. Both structures are quite similar. The amino terminus is on the left side of the figure.

cellular contents.[40] It is one of the simplest known proteins and is unique in that it folds correctly only in nonpolar environments (bound to micelles or in nonpolar solvents). Aqueous solvents disrupt the structure.[41] The sequence is[42]

Gly-Ile-Gly-Ala-Val-Leu-Lys-Val-Leu-Thr-Thr-Gly-Leu-Pro-Ala-Leu-
Ile-Ser-Trp-Ile-Lys-Arg-Lys-Arg-Gln-Gln-$NH_2$

The preferred conformations for the two major nonpolar segments (Val-5-Leu-9 and Leu-13-Ile-20) were computed first.[9] The global minimum for both was an $\alpha$-helix although other nonhelical structures for low-energy were also obtained.[9] Minimization of the energies of thousands of starting conformations for the Val 5-Leu 9 sequence, for example, resulted in only 10 low-energy minima as shown in Table I. The lowest energy structures (conformers 1 and 2) are $\alpha$-helical, while other conformers such as conformer 10 are nonhelical. Combination of the minima obtained for this peptide segment with those for the intervening segment Thr-Thr-Gly, followed by energy minimization of the resulting conformations, resulted in a small number of predominantly $\alpha$-helical structures. These low-energy conformations were then combined with those for the Leu-13-Ile-20 sequence, repeating the procedure described above. Finally, the Gly-1-Ala-4 sequence was added, one amino acid at a time, to the resulting low-energy conformations for the Leu-5-Ile-20 sequence.[9]

Only two lowest energy structures were found, both being quite similar to each other, as shown in Figure 2. The next lowest energy structure was nearly 10 kcal/mol higher in energy.[9] Both of the structures of Figure 2 are $\alpha$-helical from Gly-1 to Thr-10 and from Pro-14 to Ile-20. These two helices are separated by a bend at Thr-11-Gly-12, the difference in the two structures occurring in this bend region.[9] The helical axes form an obtuse angle of about 110°, and the nonpolar side chains pack in the interior of the molecule causing significant stabilization of the structure while the polar residues protrude toward the outside of the molecule, as shown in Figure 2. These features are in close agreement with the results of physical studies on the structure of melittin, including the X-ray crystal structure[40,43] of a melittin tetramer (which, because the monomer units are highly positively charged, tended to have a somewhat wider angle between helices so as to minimize charge repulsion; however, the positions of the helices, the bend, and the packing arrangement of side chains were essentially identical with the same features of the calculated structure).

Recent experimental work on the mechanism of action of melittin suggests that the conformation of

**Table I.**
**Low-Energy Conformations[a] of *N*-Acetyl-Val-Leu-Lys-Val-Leu-$NHCH_3$[9]**

| | conformational state[b] | | | | | rel energy, |
|---|---|---|---|---|---|---|
| conformer | Val | Leu | Lys | Val | Leu | kcal/mol |
| 1 | A | A | A | A | A | 0.0 |
| 2 | A | A | A | A | C | 0.3 |
| 3 | C | A* | G | A | C | 1.4 |
| 4 | A | C | G | A | A | 1.4 |
| 5 | C | A* | G | A | C | 1.9 |
| 6 | A | C | G | A | C | 2.1 |
| 7 | A | A | A | C | C | 2.5 |
| 8 | C | A | C | A | D | 2.6 |
| 9 | C | D | D | C | A | 2.7 |
| 10 | C | D | D | C | D | 2.8 |

[a] With energies within 3 kcal/mol of that of the global minimum (conformer 1). [b] Conformational states are defined as follows: States in which the dihedral angles are represented by a single-letter code are defined in ref 16. The familiar states are A ($\alpha$-helix) and E (extended). The actual dihedral angle ranges for all single-letter states are as follows: A, $-110° \leq \phi < -40°$, $-90° < \psi < -10°$; C, $-110° < \phi < -40°$, $50° < \psi < 130°$; D, $-180° < \phi < -110°$, $20° < \psi < 110°$; E, $-180° < \phi < -110°$, $-180° < \psi < -140°$ or $110° < \psi < 180°$; F, $-110° < \phi < -40°$, $-180° < \psi < -140°$ or $130° < \psi < 180°$; G, $-180° < \phi < -110°$, $-90° < \psi < -40°$. States indicated by asterisks are obtained by multiplying the corresponding single-letter-state dihedral angles by −1 and reversing the inequalities. All energies are relative to the energy of conformer 1. Where conformational states are identical, side-chain rotameric states differ.

(40) T. C. Terwilliger, L. Weissman, and D. Eisenberg, *Biophys. J.*, **37**, 353–361 (1982).
(41) L. R. Brown, W. Braun, A. Kumar, and K. Wüthrich, *Biophys. J.*, **37**, 319–328 (1982).
(42) E. Habermann and J. Jentch, *Hoppe Seyler's Z. Physiol. Chem.*, **348**, 37–50 (1967).
(43) T. C. Terwilliger and D. Eisenberg, *J. Biol. Chem.*, **257**, 6016–6022 (1982).

**Table II.**
**Low-Energy Conformations[a] for the Nonpolar Decapeptide (Residues 6–15) in the "Normal" P21 Protein *N*-Acetyl-Leu-(Val)$_3$-Gly-Ala-Gly-Gly-Val-Gly-NHCH$_3$[10]**

| | conformational state[b] | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| conformer[b] | Leu | Val | Val | Val | Gly | Ala | Gly-12 | Gly | Val | Gly | energy, kcal/mol |
| 1 | A | A | A | A | A | C | D* | A | A | C | 0.0 |
| 2 | A | A | A | A | A | A | C | D* | A | A | 0.8 |
| 3 | A | A | A | E | D* | D | D* | A | C | F* | 0.9 |
| 4 | A | A | A | A | A | D | D | D* | A | A* | 1.2 |
| 5 | A | A | A | A | A | A | C | D* | A | A* | 1.2 |
| 6 | A | A | A | A | A | C | D* | A | A | A* | 1.6 |
| 7 | A | A | A | E | D* | D | D* | A | C | D* | 1.7 |
| 8 | A | A | A | A | A | C | D* | A | A | D* | 1.7 |
| 9 | A | A | A | A | A | C | D* | A | A | A | 1.9 |
| 10 | A | A | A | A | A | A | C | D* | A | D | 1.9 |
| 11 | A | A | A | E | D* | D | D* | A | C | E* | 2.0 |
| 12 | A | A | A | A | A | A | A | A | A | A | 2.0 |

[a] Low-energy conformations are defined as those whose energies lie within 2 kcal/mol of that of the global minimum (conformer 1). [b] See footnote *b* of Table I for definition of conformational states.

melittin may change in the presence of a trans-membrane potential.[44] In the presence of a trans-membrane potential of approximately −90 mV, it appears that the molecule elongates so that it protrudes through both sides of the membrane, thereby forming a pore or channel. In the absence of such a voltage, it appears to adopt a more compact form that does not extend through both sides of the membrane.[44] The electrical energy provided by the trans-membrane voltage would be sufficient to increase significantly the fraction of conformers of melittin that can exist in a more open, less compact form,[44] very likely in the bend region, Thr-11-Gly-12-Leu-13.[9,43,44]

Both the leader sequence described above and melittin share the common property that they each have long, contiguous sequences of nonpolar residues that interact with membranes. These sequences induce the formation of $\alpha$-helical conformations which end in kinks in the chain at such residues as Pro and Gly. The oncogene-encoded P21 proteins also contain a long nonpolar sequence, the transforming region[10] involved in malignant transformation of cells. This sequence terminates with Gly residues and is preceded by a polar sequence, reminiscent of the murine pre-$\kappa$-leader sequence discussed above. Therefore, the methods used to compute the three-dimensional structures of the simpler membrane polypeptides have also been applied to determine the preferred structures of the transforming region of the P21 protein, to gain insight into the structural basis of oncogenesis.

## Structure of the Transforming Region of the Oncogene-Encoded P21 Proteins

A remarkable breakthrough in cancer research has been the discovery that specific human genes (oncogenes) can be introduced into normal cells in a process called transfection and transform them into malignant cells.[3,4] The best studied of these genes is the EJ human bladder carcinoma oncogene, a member of the *ras* gene series, which was the first human gene shown to be capable of transforming normal (NIH 3T3) cells into malignant ones.[3,4] From hybridization studies, this gene was found to be almost identical in its base sequence with a gene that is present in normal cells (called a protooncogene) which, when transfected into normal cells, does *not* cause tranformation.[3,4] Sequencing of both the normal gene (protooncogene) and the malignancy-causing gene (oncogene) revealed only one base change between the two genes.[3-5] This change occurs in the twelfth coding triplet in which GGC, the codon for Gly, is changed to GTC, the codon for Val.[3-5] Both genes code for proteins which contain 189 amino acid residues and have a molecular weight of 21 000, hence the term P21 proteins. Thus, a simple change in one amino acid at position 12 in the P21 protein causes the expression of a malignant phenotype.[3-5] The level of gene expression, i.e., rate of protein synthesized in cells, for both genes is the same, indicating that the abnormality caused by the oncogene is due to the protein product itself rather than to an aberration in DNA control mechanisms. This conclusion has now been verified directly by microinjection of the oncogene-encoded P21 protein into NIH 3T3 cells which underwent transient malignant transformation as long as the abnormal protein remained in the cell.[45] These results, together with the results of conformational energy calculations discussed below, suggest that a structural change in the P21 protein is responsible for its transforming properties.

The substitution of the nonpolar Val for Gly at position 12 occurs in the middle of a nonpolar decapeptide from Leu-6 to Gly-15, called the transforming region,[10] flanked by two polar sequences. The sequence for the first 20 amino acids of the P21 protein is[3-5]

1 10
Met-Thr-Glu-Tyr-Lys-Leu-Val-Val-Val-Gly-Ala-*Gly*-Gly-Val-Gly-Lys-
20
Ser-Ala-Leu-Thr

The italicized Gly at position 12 is the residue replaced by Val in the EJ transforming protein. This segment of the protein is implicated as part of a nucleotide binding site.[46,47] The P21 proteins are known to bind GTP with high affinity.[48] The nonpolar segment (residues 6–15) appears to contain strong structural determinants and may also be associated with the cell membrane.[46] The computational methods described

(44) C. Kempf, R. D. Klausner, J. N. Weinstein, J. Van Renswoude, M. R. Pincus, and R. Blumenthal, *J. Biol. Chem.*, **257**, 2469–2476 (1982).

(45) D. W. Stacey and H.-F. Kung, *Nature (London)*, **310**, 508–511 (1984).

(46) R. K. Wierenga and W. G. J. Hol, *Nature (London)*, **302**, 842–844 (1983).

(47) N. J. Gay and J. E. Walker, *Nature (London)*, **301**, 262–264 (1983).

(48) J. P. McGrath, D. J. Capon, D. V. Goeddel, and A. D. Levinson, *Nature (London)*, **310**, 644–649 (1984).

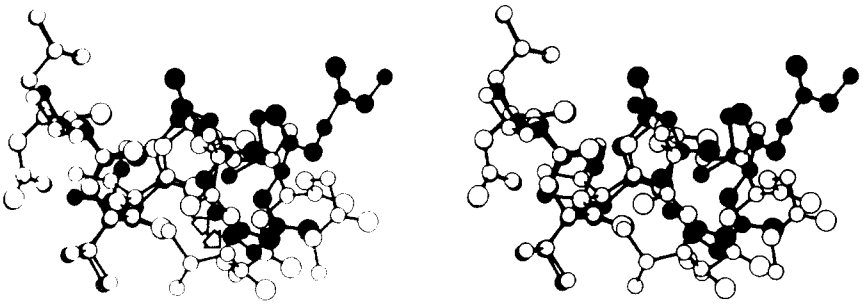


**Figure 3.** Stereoview of the superposition of the oncogenic V peptide (open circles) on the normal G peptide (solid circles). The arrows point to the $C^{\alpha}$'s of residues 12 where the two structures begin to deviate from one another. The amino terminus is on the left side of the figure.[10]

**Table III.**
**Low-Energy Conformations[a] for the Nonpolar Decapeptide (Residues 6–15) in the Transforming P21 Protein *N*-Acetyl-Leu-$(Val)_3$-Gly-Ala-Val-Gly-Val-Gly-$NHCH_3$[10]**

| | conformational state[b] | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| conformer | Leu | Val | Val | Val | Gly | Ala | Val-12 | Gly | Val | Gly | energy, kcal/mol |
| 1 | A | A | A | A | A | A | C | D* | A | C | 0.0 |
| 2 | A | A | A | A | A | A | C | D* | A | A* | 1.0 |
| 3 | A | A | A | A | A | A | C | D* | A | D* | 1.9 |
| 4 | A | A | A | C | D* | A | A | A | A | A | 2.0 |

[a] Low-energy conformations are defined as in footnote *a* of Table II. [b] See footnote *b* of Table I for definition of conformational states.

above have thus been applied to the nonpolar segment.

The results[10] for the decapeptide from the normal protein with Gly at position 12 (herein referred to as the G peptide) are shown in Table II. Numerous conformations are available to this peptide, many of which, such as conformer 3, contain significantly non-$\alpha$-helical structures. In fact, the all-$\alpha$-helical structure (conformer 12) is over 2 kcal/mol higher in energy than the global minimum (conformer 1). This global-minimum structure (Figure 3) contains an $\alpha$-helix from Leu-6 to Gly-10 that terminates in a distinct bend conformation at Ala-11 and Gly-12. In this bend structure, which also exists at residues 11 and 12 in many of the structures in Table II, Gly adopts the D* state, a left-handed twist structure that is energetically forbidden for L-amino acids.[15-17]

Table III summarizes the results[10] for the same nonpolar P21 peptide with Val at position 12 (herein referred to as the V-peptide) as occurs in the EJ bladder carcinoma P21 protein. There are many fewer minima, and these are distinctly more $\alpha$-helical than those found for the G-peptide. The lowest energy structure (conformer 1) contains an $\alpha$-helix that extends from Leu-6 to Ala-11 and terminates in a bend (CD*) at Val-12-Gly-13; i.e., the substitution of Val for Gly at position 12 lengthens the helix by one residue, to Ala-11. This structure is identical with the second lowest energy structure for the G-peptide (conformer 2, Table II). Of course, no minima exist for the V-peptide that are identical with the global minimum for the G-peptide (conformer 1, Table II) because Val cannot adopt the D* state.[10,15-17] The two global minima for the G- (solid circles) and V- (open circles) peptides are compared in stereoview in Figure 3.[10] It may be seen that the two structures are quite similar to one another up to the arrow ($C^{\alpha}$'s of residues 12) where the two structures diverge significantly.[10]

These results suggest that the G-peptide can adopt a unique conformation (or set of conformations) with residues 11 and 12 in the CD* conformation which no 6–15 P21 peptide with any L-amino acid substituted for Gly-12 can adopt. Thus, substitution of L-amino acids at position 12 would be expected to cause an abnormally functioning protein.[10,49-51] This conclusion has been corroborated by the isolation of other naturally occurring P21 transforming proteins that are identical with the EJ P21 protein except that they contain Ser,[3,4] Lys,[3,4] or Asp[52] at position 12. These represent arbitrary substitutions, yet all are transforming. Quite recently, site-specific mutations have been introduced into the 12th coding triplet in the cloned human *ras* oncogene so that every one of the naturally occurring amino acids has been introduced into position 12 of the P21 protein.[53] All except Gly and Pro have been found to be transforming.

The fact that the G-peptide can adopt a low-energy conformation (conformer 2 Table II) which is identical with the global minimum for the V-peptide suggests that, if the normal P21 protein is present at elevated intracellular concentrations at which (according to the Boltzmann principle) this alternate structure can be present in significant amounts, it may exhibit transforming activity.[10,49-51] A transfection experiment was performed with the protooncogene that was spliced onto a viral LTR (long terminal repeat sequence) which allows unrestricted expression of neighboring genes.[54] This gene caused transformation of NIH 3T3 cells which produced high amounts of normal P21 protein.[54] Thus, two major results deduced from a comparison of the structures of the G- and V-peptides have been corroborated by genetic experiments.[10,49-51]

(49) M. R. Pincus in "Molecular Basis of Cancer, Part A: Macromolecular Structure, Carcinogens, and Oncogenes", R. Rein, Ed., Alan R. Liss Inc., New York, 1985, pp 419–430.
(50) M. R. Pincus and P. Brandt-Rauf, Eighteenth Middle Atlantic Regional Meeting of the American Chemical Society, Proceedings, **40**, (1984).
(51) M. R. Pincus and P. W. Brandt-Rauf, *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 3596–3600 (1985).
(52) E. Santos, E. P. Reddy, S. Pulciani, R. J. Feldmann, and M. Barbacid, *Proc. Natl. Acad. Sci. U.S.A.*, **80**, 4679–4683 (1983).
(53) P. H. Seeberg, W. W. Colby, D. J. Capon, D. V. Goeddel, and A. D. Levinson, *Nature (London)*, **312**, 71–75 (1984).
(54) E. H. Chang, M. E. Furth, E. M. Scolnick, and D. R. Lowy, *Nature (London)*, **297**, 479–483 (1982).

Since arbitrary substitutions of L-amino acids (except for Pro) for Gly at position 12 of the P21 protein all result in transforming activity, the preferred conformations induced by the presence of other L-amino acids at position 12, particularly Ser, Lys, and Pro were explored[49-51] to determine if there were any common structural features among the abnormal proteins. The global minima (not shown here) for the Lys-12-containing peptide (K-peptide) and the Ser-12-containing one (S-peptide) are identical with each other and to the global minimum for the V-peptide.[49-51] Furthermore, regardless of the substitution (Val, Lys, or Ser), the peptides remain predominantly $\alpha$-helical. None of these peptides (V, K, and S) have helix termination at residues 11 and 12 as in the G-peptide.[10,49-51] These findings suggest the existence of a "malignancy-causing" conformation that has helix propagation through Ala-11.

Support for such a conformation is provided by the results obtained from the substitution of Pro for Gly at position 12 of the decapeptide. (This peptide with Pro-12 is herein referred to as the P-peptide.) As in the case of the G-peptide, numerous nonhelical conformations were obtained for the P-peptide. The global minimum contains a distinct break in the helix at Ala-11 and, in a second conformer of almost equal energy, a break in the helix at Ala-11 and Pro-12.[49-51] No structure contains helix propagation to (but not including) residue 12 as occurs in all of the other peptides with L-amino acids at position 12.[49-51] This result is to be expected since residues preceding Pro cannot adopt the A ($\alpha$-helical) conformation without bad contacts between several of the backbone atoms of Pro and the residue preceding it.[9,18] Thus, helix termination occurs at the residue preceding Pro (as in the murine pre-$\kappa$-leader sequence and in melittin) which is Ala-11 in the P21 protein. The G- and P-peptides share the common feature that helix termination occurs at Ala-11, a feature not shared by the P21 peptides with other L-amino acids at position 12. Because it does not adopt the putative "malignancy-causing" conformation (conformer 1 of Table III) but rather adopts conformations similar to the global minimum for the G-peptide (conformer 1, Table II), the P-peptide-containing protein would be expected to have low transforming ability as in fact is found experimentally,[53] as discussed above. Since Ala 11 is less likely to be $\alpha$-helical in the P-peptide than in the G-peptide, the P-peptide would be expected to have an even lower transforming potential.

Thus, application of conformational energy calculations to the prediction of the structures of P21 peptides yields results that can be tested in genetic experiments and provides a structural basis for accounting for particular experimental genetic findings.

## Conclusions

Conformational energy calculations have been used to compute single-residue minima for amino acid residues and simple di-, tri-, and oligopeptides.[6,7] These methods have now been extended to polypeptides and simple proteins and have enabled us to compute the most likely three-dimensional structures for these peptides which are in agreement with experimental data. Computation of these structures allows evaluation of the interactions that stabilize particular conformations, of the energy change necessary to alter the conformation of a peptide from one state to another (as in the case of the P21 peptides), and of the structural effects of amino acid substitutions in given polypeptide chains.

The recent results on the different peptides that we have presented here illustrate several important general principles of protein folding that can be used in applying the methodology described here to determine the structures of large proteins. First, as has been discussed at length elsewhere,[6,7] short-range interactions in polypetides and proteins suffice to give all of the structures of di- and tripeptides that they will eventually adopt in the final folded form of the protein. Subsets of these conformations then become "selected out" as more long-range or medium-range interactions (such as the $i \rightarrow i + 4$ hydrogen bond in an $\alpha$-helix) are brought into play. From the successive incorporation of the latter interactions, as occurs in the $\beta$-sheet conformation of gramicidin S and in the $\alpha$-helices of melittin, fewer conformations remain viable as the chain length is increased. Finally, long-range interactions such as the packing of nonpolar residues in melittin and the side chain–backbone interaction in gramicidin S begin to contribute toward "selection" of the lowest energy structure(s). An important principle of chain-folding initiation also emerges from these calculations. Specific sequences such as Val-5-Leu-9 in melittin show strong conformational preferences and cause neighboring sequences such as the relatively unstructured sequence Thr-Thr-Gly to adopt compatible conformations. Thus, identification of CFIS's in proteins is important for applying this method to large proteins. Currently, experimental methods for identifying CFIS's in proteins are being developed.[36,55,56]

Another type of chain-folding-initiation phenomenon encountered in these calculations was long-range CFI as occurred in the case of collagen, where, of many possible poly(Gly-Pro-Pro) conformations for the single chain, only one was selected by the specific interactions among three single chains. Thus, it is likely that the association of two or more nonadjacent segments in a protein can lead to strong favorable interactions that are highly specific and greatly limit the allowed conformations for the whole polypeptide chain.

*This work has been supported by the National Foundation for Cancer Research, by the National Institute of General Medical Sciences (GM-14312), and by the National Science Foundation (DMB84-01811).*

(55) L. G. Chavez, Jr., and H. A. Scheraga, *Biochemistry*, **16**, 1849–1856 (1977).

(56) R. P. Carty, F. Gerewitz, and M. R. Pincus, *Biophys. J.*, **37**, 95A (1982).