An Application of Constrained Optimization in Protein Folding: The Poly-L-Alanine Hypothesis

David M. Gay, Teresa Head-Gordon, Frank H. Stillinger and Margaret H. Wright,
AT&T Bell Laboratories, Murray Hill, New Jersey

Introduction

The protein folding problem is recognized as one of the grand challenges in the biophysical sciences today, and has been surveyed in several recent articles (see, e.g., Richards, 1991). Many approaches to this problem, some involving substantial numerical computation, have been suggested over the past three decades. Much of the pioneering work has been carried out at Cornell University by Harold Scheraga and collaborators, and Cornell remains an active center of leading-edge research.

The mainstream optimization community, particularly researchers in global optimization, have begun to take a serious interest in the protein folding problem, as evidenced by two related mini-symposia at the Society of Industrial and Applied Mathematics (SIAM) Optimization Conference held in May 1992. The relevance of protein folding to optimization arises from the widely held theory that the desired ("native") structure of a given protein corresponds to a minimizer (perhaps the global minimizer) of an empirical potential energy function. Although this might appear to suggest that the native structure can be found simply by invoking a "black box" unconstrained minimization package, such a naive approach is impossible because of the size and complexity of the associated optimization problems. Not only is the number of variables very large even for proteins of modest size—for example, BPTI (bovine pancreatic trypsin inhibitor) corresponds to 1053 variables—but also the number of distinct local minima of the standard potential energy function increases exponentially with the number of variables.

The authors of this article (two chemists and two optimizers) have recently worked on the "poly-L-alanine hypothesis," which may provide some insights for the protein folding problem. The idea is that there is a close correspondence between the backbone geometry of the native-structure energy minimum for any polypeptide or protein of M residues, and that of a local minimum of poly-L-alanine with the same number of residues. Validation of this hypothesis can lead to several conceptual advantages. First, although mechanical stability of a native structure does not depend on side-chain details, free-energy stability is controlled by those details; recent work (Zhang *et al.*, 1991) shows that replacement of several residues by L-alanine preserves native structure. Second, poly-L-alanine can serve as a natural reference material in theoretical calculations of the relative stabilities of distinct folded structures for a given protein. A third consequence is that the specific structural roles of disulfide bonds, charged side chains, packing of amino acid side chains, and special residues can be assessed quantitatively. Finally, new experiments and computer simulation studies are suggested.

The calculations performed in our work involved polypeptides ranging in size from very small to reasonably large. The results, which are reported in detail in Head-Gordon *et al.* (1991), showed that poly-L-alanine can adopt an impressive number of structural alternatives. Here we discuss only a single topic: the application of a method for nonlinearly constrained optimization.

Problem formulation

The empirical potential energy function used in our study has the form described in Brooks *et al.* (1983), namely a sum of five highly nonlinear terms. Four terms of the sum refer to chemical bond connectivity. The fifth term represents nonbonded terms as a sum of pairwise coulomb electrostatic and Lennard-Jones interactions.

The variables are the Euclidean coordinates of the atoms; hence the number of variables is $3m$ for a protein with m atoms. Five degrees of freedom can always be eliminated, e.g. by assuming that the first atom is located at the origin, and that the x and y coordinates of the second atom are fixed at their initial values. Whether eliminating these degrees of freedom is worthwhile depends on the problem and computational method.

To support the poly-L-alanine hypothesis, it is necessary to demonstrate the existence of a local unconstrained minimizer on the poly-L-alanine hypersurface that closely mimics the native structure of a given protein. For the small peptides Ac-$(\text{ala})_n$-NHMe with $n = 3$ and $n = 8$, our approach to finding such a local minimizer was based on specifying *nonlinear constraints* that characterize the desired secondary structure—helix, turn, and sheet.

Several types of constraints were imposed as appropriate for each secondary structure: an L-chirality dihedral constraint, a peptide torsion constraint (trans or cis), backbone dihedral angle constraints (both ϕ and Ψ), and hydrogen bond constraints. Each constraint was formulated with upper and lower bounds limiting the variation of the given quantity from a target value. For the L-chirality dihedral angle, the target value was 33°, with maximum allowed variation of 3°; for peptide torsion constraints, the target was either 0° or 180°, with permitted variation of 10°;

the backbone dihedral angles ϕ and Ψ were required to lie within 0.5° of their known secondary structure values; and the hydrogen bonds were required to fall within 0.4Å of the target value 1.9Å.

For n=3, the number of variables is 126, with 13, 14, or 15 nonlinear constraints, depending on the secondary structure. For n=8, there are 276 variables, and the number of nonlinear constraints lies between 21 and 40.

Solution technique

The optimization problems were solved numerically using the code NPSOL (Gill *et al.*, 1986), an implementation of a dense sequential quadratic programming (SQP) method (see, e.g., Fletcher, 1987, for a discussion of these methods). In a generic SQP method, the search direction is the solution of a quadratic programming (QP) subproblem. Each such subproblem is itself an optimization problem of minimizing a quadratic objective function subject to linear constraints. The SQP subproblem objective function is a quadratic approximation to the Lagrangian function, and the subproblem constraints are local linearizations of the nonlinear constraints. In NPSOL, reduction in an augmented Lagrangian merit function is required at each iteration to encourage progress toward the solution.

NPSOL develops a dense positive-definite quasi-Newton approximation to the Hessian matrix of the Lagrangian function by applying a modified BFGS update that reflects changes in the gradient. The default option in NPSOL is to take the initial Hessian matrix as the identity; NPSOL also allows an initial "warm-start" Hessian to be specified by the user.

We invoked NPSOL twice for each structure, first to solve the nonlinearly con-

continued on next page

strained problem described above. The initial point for the constrained minimization was based on a structure closely resembling the desired secondary structure. The initial Hessian for the first minimization was taken as either the identity or a finite-difference Hessian of the potential energy function. Since NPSOL requires a positive-definite Hessian for the QP subproblem, the finite-difference Hessian was modified in case of indefiniteness using the Cholesky factorization. Complete details concerning various strategies for the optimization methods are given in Gay *et al.* (1992).

After achieving convergence for the constrained problem, an *unconstrained* minimization of the potential energy function was performed. For the second minimization, the initial point was the just-calculated constrained solution, and a warm-start Hessian approximation was taken as either the final quasi-Newton approximation from the constrained problem, or a (possibly modified) finite-difference Hessian. Either form of warm-start strategy had two benefits compared to the default choice of the identity: it led to substantial improvements in solution speed for the unconstrained subproblems compared to using the identity matrix, and also tended strongly to encourage convergence to a "nearby" unconstrained minimizer, as desired.

Results

Our computational experiments demonstrated the existence of local minima on the poly-L-alanine hypersurface corresponding to the usual types of secondary structure: helix, turn, and sheet. Existence of these minima implies that a variety of native sequences show stable secondary structure.

In the helix category (Creighton, 1984), the energy model was able to distinguish between the 3_{10} and α_R helix. No π helix was observed, but no experimental evidence suggests that it should be. The existence of polyproline I and polyglycine II was also demonstrated. For turns, types I, V and V' were shown to be stable minima on the alanine gas phase hypersurface for n=8. Finally, for n=8 we constructed an anti-parallel β-sheet that is a local minimizer on the poly-L-alanine hypersurface.

Although the proteins discussed are very small, the optimization problems are already challenging, since they involve not only a reasonably large number of variables, but also highly nonlinear functions. Our results indicate the likely efficacy of further research concerning numerous optimization-related topics, such as the effects of the initial Hessian approximation, tradeoffs between the gains from inclusion of second-order information and the extra work required to obtain it, possible application of "sparsified" Hessian matrices or specially structured preconditioners, and selection of termination criteria. Because of the relatively simple analytic mathematical forms of the objective and constraints, symbolic modeling languages offer great promise, both for convenient expression of larger problems and application of automatic differentiation to obtain an exact initial Hessian.

References

1. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983), CHARMM: a program for macromolecular energy, minimization, and dynamics calculations, *Journal of Computational Chemistry* 4, 187–217.

2. T. E. Creighton (1984), *Proteins, Structures and Molecular Properties*, W. H. Freeman, New York.

3. R. Fletcher (1987), *Practical Methods of Optimization* (second edition), John Wiley and Sons, Chichester.

4. D. M. Gay, T. Head-Gordon, F. H. Stillinger and M. H. Wright (1992), Use of constraints and other strategies in protein folding, manuscript, AT&T Bell Laboratories, to appear.

5. P. E. Gill, W. Murray, M. A. Saunders and M. H. Wright (1986), User's guide for NPSOL: a Fortran package for nonlinear programming, Report SOL 86-2, Department of Operations Research, Stanford University, Stanford, California.

6. T. Head-Gordon, F. H. Stillinger, M. H. Wright and D. M. Gay (1991), Poly-L-alanine as a reference material for protein folding, Manuscript, AT&T Bell Laboratories, Murray Hill, New Jersey. To appear in *Proceedings of the National Academy of Sciences*, 1992.

7. F. M. Richards (1991), The protein folding problem, *Scientific American*, 54–63.

8. X.-J. Zhang, W. A. Baase and B. W. Matthews (1991), Toward a simplification of the protein folding problem: a stabilizing polyalanine a-helix engineered in T4 lysozyme, *Biochemistry* 30, 2012–2017. ■