AFRL-SR-BL-TR-98-
0212

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE Feb. 20, 1998	3. REPORT TYPE AND DATES COVERED Final Technical, Jan. 1 94 - Dec. 31, 97

4. TITLE AND SUBTITLE

Effective energy methods for global optimization for biopolymer structure prediction

5. FUNDING NUMBERS

AFOSR
F49620-94-1-0123

6. AUTHOR(S)

Professor David Shalloway

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)

Cornell University
Ithaca, NY 14853

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

19980311 092

13. ABSTRACT *(Maximum 200 words)*

We developed the fundamental theory and algorithms of the new "Packet Annealing Method" for analyzing biopolymers 3-dimensional structures and tested it on small systems. We showed that the method provides a natural and powerful computational approximation to the stochastic description of biopolymer motions and encompasses other competing "potential smoothing" methods as special cases. Its main strength is that it uncovers and exploits the intrinsic "hidden structures" of biopolymer energy landscapes to efficiently perform global minimization using a hierarchical search procedure which concentrates parallel computing effort on a sequence of selected regions of decreasing size. Each search region corresponds to a metastable macrostate of the system, a region of conformation space that is isolated from the remainder of the space by effective energy barriers. The effective energy includes both energetic contributions from the energy potential function and entropic contributions resulting from thermal fluctuations of the biopolymer. It determines the thermodynamic macrostate free-energies which (rather than the energies) determine biopolymer structures. In addition, new methods for computing macromolecular conformational transitions and for molecular dynamics simulation were developed.

14. SUBJECT TERMS

macromolecules, biopolymers, global optimization

15. NUMBER IF PAGES

34

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Effective Energy Methods for Global Optimization for Biopolymer Structure Prediction

AFOSR G49620-94-1-0123

Final Technical Report

Feb. 20, 1998

David Shalloway

Biophysics Program

Section of Biochemistry, Molecular and Cell Biology

Cornell University

Ithaca, NY 14853.

Abstract

The ability to predict the 3-dimensional structures of biopolymers is important to AFOSR research. Current research applications include the design of biopolymer support matrices for non-linear optical materials for laser-resistant optical systems. The special characteristics of the biopolymer structure determination problem (where covalent bonding patterns are fixed) distinguish it from chemical structure determination problems (where bond patterns change) and special methods for very high dimensionality minimization are needed. Fortunately, most important problems require simpler *perturbative* global minimization and only need to be able to predicting changes in 3-dimensional conformations from a known initial conformation. This problem will be much easier to solve than the full "protein folding problem", but is still complex because of the large number of dimensions ($\sim 10^4$) involved.

We have developed the fundamental theory and algorithms of the new *Packet Annealing Method* and tested it on small systems. We showed that the method provides a natural and powerful computational approximation to the stochastic description of biopolymer motions and encompasses other competing "potential smoothing" methods as special cases. Its main strength is that it uncovers and exploits the intrinsic "hidden structures" of biopolymer energy landscapes to efficiently perform global minimization using a hierarchical search procedure which concentrates parallel computing effort on a sequence of selected regions of decreasing size. Each search region corresponds to a metastable *macrostate* of the system, a region of conformation space that is isolated from the remainder of the space by *effective energy* barriers. The effective energy includes both energetic contributions from the energy potential function and entropic contributions resulting from thermal fluctuations of the biopolymer. It determines the thermodynamic macrostate free-energies which (rather than the mean energies) determine biopolymer structures.

In addition, new methods for computing macromolecular conformational transitions and for molecular dynamics simulation were developed.

Contents

List of Figures

I. INTRODUCTION

A. Motivation: AFOSR Need for Protein Structure Prediction

The ability to predict the changes in the 3-dimensional (3-D) structures of biopolymers that are induced by changes in their covalent structures is important to AFOSR research. For example, research in the Laser Hardened Optical Materials Branch of the Electromagnetic Materials Division at Wright Laboratory, Wright Patterson AFB is aimed at developing synthetic biopolymers which, because of their well-defined 3-D structures, can provide a superior support matrix for orienting light-absorbing chromophores in non-linear optical materials for laser-resistant optical systems. Dr. Ruth Pachter, working on laser resistance technology at Wright Laboratory, concludes

> *Peptide structure predictions and molecular dynamics simulations of these peptides are key in the interpretation of the results.*

and

> *...the importance of novel developments for studying large molecular systems..., especially protein folding and design, illustrate the importance for advances in new optimization techniques for determining the global energy minimum of these compounds. Such a task will enable rapid advances in designing new laser resistant materials.* (from R. Pachter, "Nano Architectures for Agile Optical Thresholds")

Another typical application includes the development of new materials with exceptional properties modelled on naturally-ocurring proteins (e.g., exceptionally strong synthetic fibers based on the structure of spider silk). The computational needs of this research are essentially identical to those encountered in many other aspects of biotechnology—for example, in the study of the interactions between viral proteins and drugs designed to interact and interfere with them. Further, because similar chemical principles are involved, these methods should be applicable to a wide variety of other polymeric structure problems as well.

Prediction of the folding of biopolymers into their stable 3-D structures (the "protein folding problem") is difficult because of the large numbers of atomic coordinates (and hence, mathematical degrees of freedom) to be determined. Typical problems involve $10^3 - 10^4$ degrees of freedom. Fortunately, in practice it is not necessary to predict structures *de novo*. New biopolymers are experimentally developed by iterations of a design cycle in which the covalent structure of a biopolymer having a 3-D structure with fairly good characteristics is modified to a new covalent structure which, according to computational predictions, will have a modified 3-D structure with even better properties. Typical alterations include amino acid substitution or the addition, by covalent or non-covalent linkage, of a small chemical group (e.g., a chromophore) to a biopolymer of known structure. Even when determining the structure of a new biopolymer, good approximate starting points can often be selected from extensive databases of known 3-D structures (which have been experimentally determined by X-ray crystallography or nuclear magnetic resonance). In all these cases it is the simpler *perturbative structure prediction* problem that is of primary interest.

B. Global optimization, free-energies, and biopolymer structure prediction

The structure of a biopolymer is governed by its potential energy function $V(R)$, a complicated function of all the atomic coordinates $R \equiv \{\vec{r}_i; i = 1 \ldots N\}$, where N is the number of atoms. In principle, it must be derived from quantum mechanics, but since biopolymer 3-D structures are governed by non-covalent bonding (covalent bonding is invariant in most problems), classical approximations appear to be adequate. However, the very high dimensionality of the problem [N is typically $\sim O(10^4)$] makes the structure prediction problem particularly difficult and, with current algorithms, many orders-of-magnitude beyond the capabilities of even the largest parallel computers. New algorithms are needed.

Simulated annealing,[1] in which the biopolymer thermal fluctuations are simulated at a sequence of decreasing temperatures, is one of the most powerful approaches available at the present time. Although simulated annealing is inadequate for biopolymer structure prediction, it has demonstrated the value of annealing approaches in general. However, most current methods attempt to predict structure by searching for the global minimum of V, arguing that this corresponds to the most energetically stable conformation. Such purely energetic approaches ignore the entropic effects that result from conformational fluctuations and will only be accurate at very low temperatures near absolute zero, not at the working temperatures of practical importance. At working temperatures, biopolymers thermally fluctuate through many conformations according to the Gibbs-Boltzmann probability density p_B:

$$p_B(\beta; R) \propto e^{-\beta V(R)}/Z(\beta)$$
$$\beta \equiv (k_B T)^{-1}$$

where k_B is Boltzmann's constant, T is the temperature, β is the "inverse temperature", and $Z(\beta)$ is a normalizing constant (the "partition function," see Ref. 2 for review). During the course of these rapid fluctuations, the protein rapidly traverses hundreds or thousands of local minima[3] of V within an extended region of conformation space that we call a *macrostate*. Over longer time periods, particularly at higher temperatures, the biopolymer will occasionally make transitions to other extended macrostate regions. The probability of being in each macrostate is given by its *free-energy*, which is the integral of $p_B(\beta; R)$ over the macrostate region[2] and which depends both on V within the macrostate and on the size of the macrostate. Thus, it is the free-energy that must be globally minimized during annealing to predict biopolymer structure.

Each macrostate is separated from the others by energy barriers that must be large compared to the thermal energy $k_B T$, so conformational fluctuations *within* a macrostate are *rapid* while conformational fluctuations *between* macrostates are *slow*. Furthermore, the size and free-energy of each macrostate depends on the temperature. Even the number of macrostates varies with temperature: small "child" macrostates that exist at low temperatures will merge into unified "parent" macrostates at higher temperatures when the energy barriers between them become small compared to $k_B T$. In principle, the properties of the individual macrostates (e.g., mean conformation, enthalpy, entropy, etc.) and the connections between them can be computed and traced in *macrostate trajectory diagrams*. These relationships and diagrams can provide a hierarchical description of conformation space that reflects the underlying kinetic properties of the biopolymer. The macrostates constitute a "hidden structure"[4] that strongly influences algorithmic performance and can be used to advantage once uncovered.

C. A physical analogy

To illustrate, consider the 2-dimensional problem of finding the lowest region on the surface of the earth. Simulated annealing corresponds to tracking the position of a very small test-object as it migrates while the earth is shaken with progressively lower intensities (temperatures). The process is inefficient because after each jump the test-object samples the height (energy) over a region that is too small compared with the sizes of the jumps. Because of this, the test-object is too sensitive to small-scale local fluctuations in the fractal-like energy surface that tend to mask the more important large-scale global structures of the surface. A more efficient procedure would be to start with large "soft" test-objects with diameters matched to the sizes of the oceans (e.g., a 10,000 km beach-ball) and to iteratively minimize their positions as temperature was progressively reduced. The sizes of the beach-balls should be matched to the landscape in a self-consistent manner so that the balls are roughly of the same size as the temperature-dependent confining regions that they are searching, i.e., the macrostates of the system. The search trajectory of each ball will be governed by the *effective energy*, an integral of the height over a region self-consistently chosen to match the size of the ball. At high temperature the macrostates are the oceans. As temperature (shaking) is decreased, the ridges separating smaller depressions in the bottom of the oceans become important and the oceanic macrostates divide (*branch*) into smaller child macrostates. For an efficient parallel search, a separate ball should initially be used to search each ocean and, as temperature is reduced, each ball should be replaced with an appropriate set of smaller balls to search each child macrostate region. This process continues recursively as children have children and the macrostates get smaller and smaller. Since we expect their number to increase rapidly with decreasing temperature, all macrostates can not be searched and it is necessary to select only the most promising for investigation. Success requires that the number of macrostates does not grow rapidly in comparison with our ability to discard unprofitable search trajectories. We call this approach the *Packet Annealing Method* (see Church et al., 1996).

The Packet Annealing Method is particularly suited to the fractal-like structure of the surface of the Earth: because it has been formed by the action of a large number of quasi-independent forces, the surface contains structure at multiple spatial scales. Biopolymer energy functions probably have similar properties since they are sums of very large numbers of quasi-independent interactions dominated by two-body terms. Note that this algorithm will not find anomalous minima–deep but very narrow holes (e.g., oil wells). It is the fact that the algorithm is explicitly designed to ignore such anomalous minima that makes it highly efficient. This is not a disadvantage since anomalous minima of the energy function are not usually minima of the free-energy at practical temperatures and, in any case, would not be expected to be found by the physical system itself.

D. Packet Annealing Method

During the project we developed most of the formalism and algorithms needed to implement the Packet Annealing Method. The central tool is the effective energy:

$$\widetilde{H}_K(\beta;R) \equiv -2\beta^{-1}\log\left[\det^{-\frac{1}{4}}\left(\frac{\pi}{\beta K}\right)\int e^{-\frac{\beta}{2}V(R')}\,e^{-\frac{\beta}{2}(R'-R)K(R'-R)}\,dR'\right] \qquad (1)$$

which depends on the integral of $p_B^{\frac{1}{2}}(\beta;R)$ over a Gaussian-weighted region parameterized by a fluctuation tensor $\Lambda \equiv (2\beta K)^{-\frac{1}{2}}$. As $\Lambda \to 0$, $\widetilde{H}_K$ reduces to V. For non-zero Λ, $\widetilde{H}_K$ is a smoothed transform of V that suppresses all fluctuations on size-scales $< \Lambda$. Because it is smoothed, $\widetilde{H}_K$ can be minimized much more rapidly than V.

We have shown (Oresic and Shalloway, 1994; Shalloway,1996) that $\widetilde{H}_{K^0_\alpha}(\beta; R^0_\alpha)$ provides a good approximation to the free-energy of macrostate α when R^0_α is set to the centroid of the macrostate and K^0_α is properly matched to the size of the macrostate. Each macrostate probability distribution can be approximated by a Gaussian *characteristic packet* ϕ^0_α:

$$\phi^0_\alpha(\beta; R) = e^{-\frac{\beta}{2}\{V^0_\alpha(\beta)+[R-R^0_\alpha(\beta)]\,K^0_\alpha(\beta)\,[R-R^0_\alpha(\beta)]\}}/Z^{\frac{1}{2}}(\beta) \tag{2}$$

which is described by the *characteristic parameters* R^0_α, K^0_α, and an amplitude-fixing scalar V^0_α. These parameters are determined by solving the

Characteristic Packet Equations

$$\begin{array}{rcll}
 & \textbf{Integral form} & \textbf{Differential form} & \\
(R^0_\alpha)_i = & \dfrac{\langle p_B^{\frac{1}{2}} | R_i | \phi^0_\alpha \rangle}{\langle p_B^{\frac{1}{2}} | \phi^0_\alpha \rangle} & \left.\dfrac{\partial \widetilde{H}_{K^0_\alpha}(\beta; R^0)}{\partial R^0_i}\right|_{R^0=R^0_\alpha} = 0 & \text{(3a)} \\
(\Lambda^0_\alpha)^2_{ij} = & \dfrac{\langle p_B^{\frac{1}{2}} | (R-R^0_\alpha)_i (R-R^0_\alpha)_j | \phi^0_\alpha \rangle}{\langle p_B^{\frac{1}{2}} | \phi^0_\alpha \rangle} & \left.\dfrac{\partial^2 \widetilde{H}_{K^0_\alpha}(\beta; R^0)}{\partial R^0_i\, \partial R^0_j}\right|_{R^0=R^0_\alpha} = (K^0_\alpha)_{ij} & \text{(3b)}
\end{array}$$

and

$$V^0_\alpha(\beta) = \widetilde{H}_{K^0_\alpha}(\beta; R^0_\alpha) - 2\beta^{-1} \log \det{}^{-\frac{1}{4}} \left(\frac{\pi}{\beta K^0_\alpha} \right) . \tag{4}$$

Eqs. (3) are a non-linear coupled set of equations that, in effect, perform pattern-recognition to identify the dominant structures of the energy landscape at each temperature. Each solution corresponds to a macrostate. For example, Fig. 1 displays p_B at a sequence of decreasing temperatures for a model two-dimensional potential V (panel a). The support of p_B is dispersed at high temperature and converges at low temperature to a small region centered about the global minimizer of V. While p_B is complicated at intermediate temperatures (panel b), the behavior of the macrostates at different temperatures can be simply modeled by following the appearance, movement and disappearance of regions of concentrated probability density with temperature. At every temperature, each region α is characterized by the characteristic parameters (panel d) and corresponding characteristic packets (panel c). As T decreases, the characteristic packets and corresponding macrostates

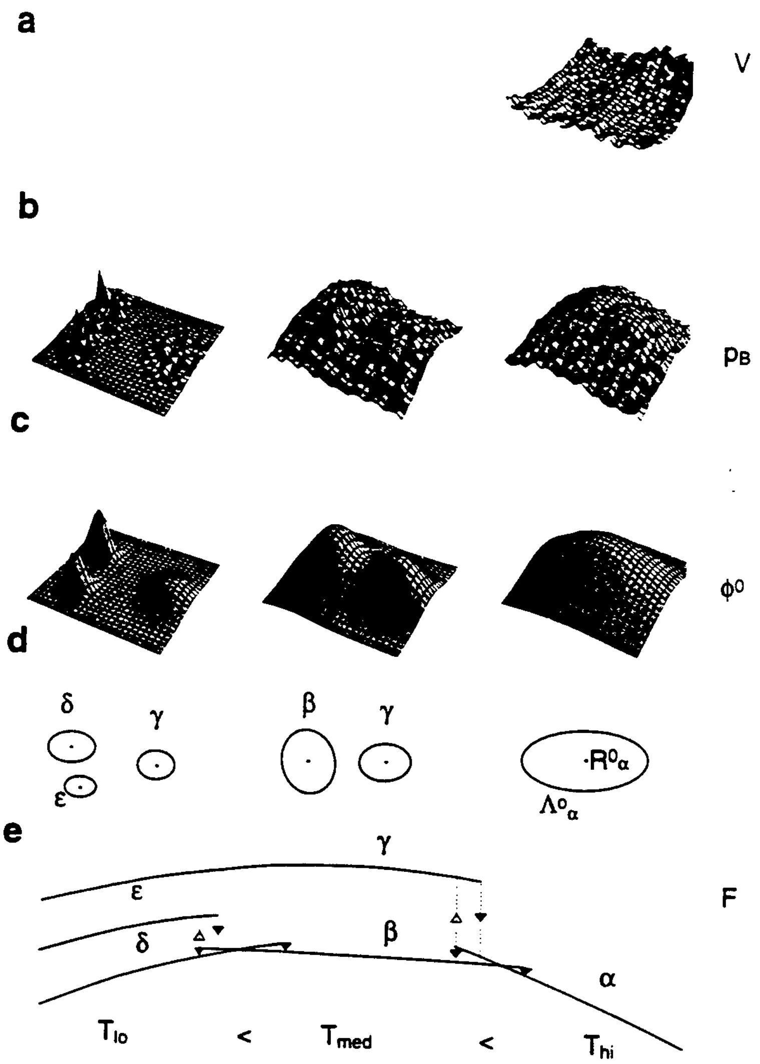

FIG. 1. Annealing using metastables states. (a) A model two-dimensional potential $V(r_1, r_2)$. (b) The corresponding Gibbs/Boltzmann probability distribution p_B at three temperatures, $T_{\mathrm{hi}} > T_{\mathrm{med}} > T_{\mathrm{lo}}$. (c) Superposition of the (squared) characteristic packets $(\phi^0_\alpha)^2$ that are solutions to the packet equations at the three temperatures. (A large number of characteristic packets, corresponding to the very small-scale fluctuations of V, will appear at lower temperatures.) (d) The characteristic packets are parametrized by the positions of their center-of-masses (R^0_α) and by their root-mean-square fluctuations tensors (Λ^0), represented here by ellipses. (e) Free-energy vs temperature trajectory diagram for this temperature range. Solid lines represent metastable state trajectories and dotted lines represent transitions. The discontinuities in the trajectories correspond to branch points at which packets bifurcate. Solid arrowheads indicate "escape" or preferred "capture" transitions. Open arrowheads indicate unpreferred transitions that can be detected by the missing-mass procedure. (From Church et al., 1996).

continuously decrease in size and divide into children. This process is reported by the trajectory diagram (panel e) which provides hierarchical description of the energy landscape which displays its intrinsic structure in a manner that is not obvious from inspection of V itself. At each temperature the most stable macrostate of the system is the one having lowest free-energy (i.e., macrostate α, β or δ, depending on T).

E. Trajectory diagrams and scaling properties

It is frequently speculated that protein energy landscapes have an overall structure that naturally "funnels" the macromolecules towards their native folded state.[5] This could explain the fact that natural proteins fold very rapidly compared to the rate that would be expected if they performed a random search for the native state.[6] However, it has not previously been possible to computationally determine whether this was true except for highly idealized model systems. The trajectory diagrams enable us, for the first time, to do this.

Consider the two potentials $V(R)$ shown in Fig. 2. The one on the left will funnel systems to the lowest energy state; the one on the right will not. This is reflected in the variation of $\widetilde{H}_K$ with Λ (plotted above): a sequence of local minimizations of $\widetilde{H}_K$ as Λ is reduced converges to the global minimum in the funneling case but not in the non-funneling case. We call the former case *strong scaling* and the latter case *weak scaling* to emphasize the role of the scaling parameter Λ.

While the scaling properties of these simple one-dimensional cases can be determined by graphical examination of V and $\widetilde{H}_K$, this is not possible in high dimensionality problems. However, the scaling properties can be determined from the trajectory diagrams in any number of dimensions. The critical point is that in the strong scaling case the trajectory that leads to the low-temperature global minimum is the lowest trajectory at all temperatures. Thus it is very easy to trace. In contrast, there is *trajectory crossing* in the weak scaling case and it is not sufficient to trace just the lowest-lying trajectory at each temperature. The weaker the scaling, the more low-lying trajectories must be traced to be certain that

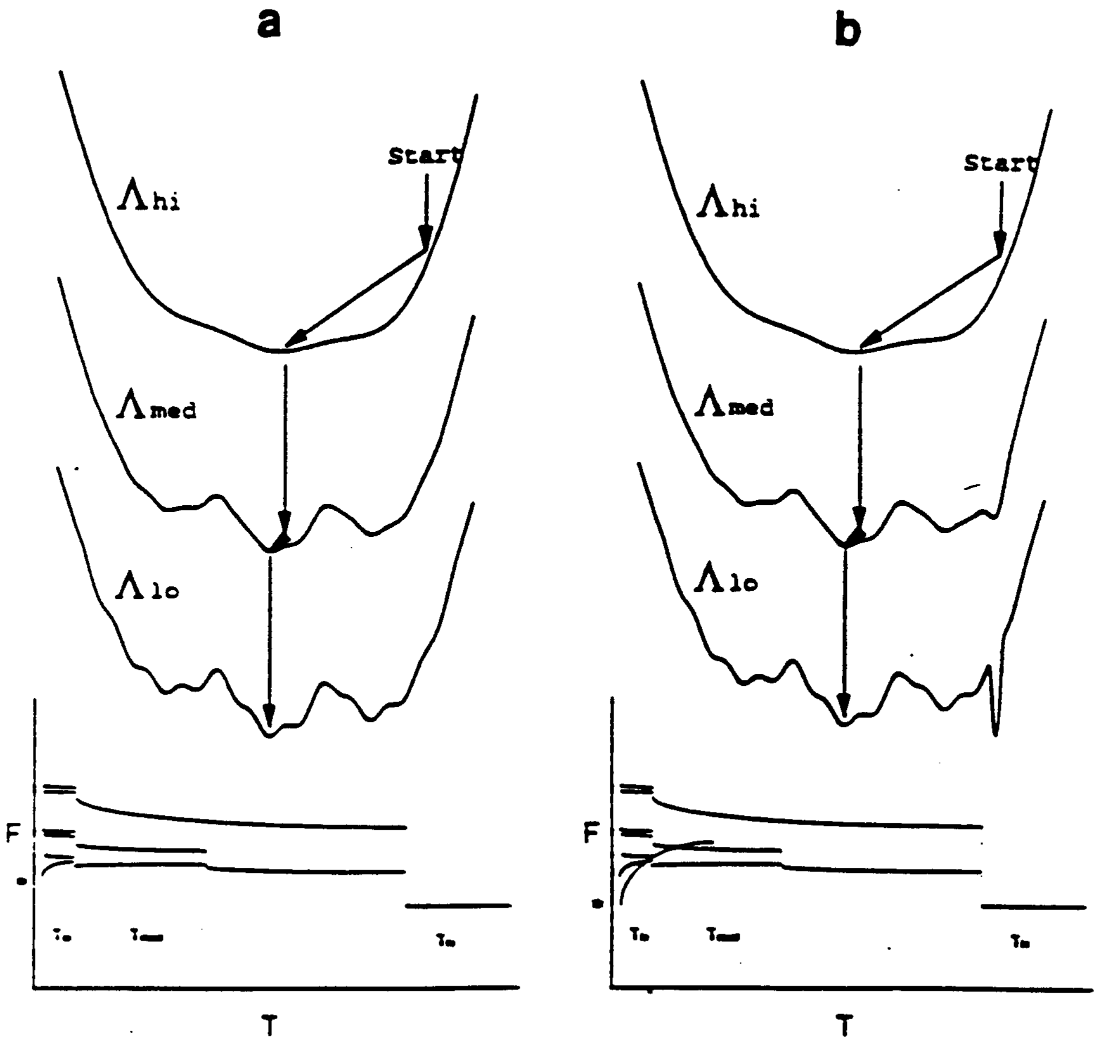

FIG. 2. Strong and weak scaling properties. Each upper panel shows a sequence of effective-energy functions obtained by convolution with Gaussians having widths $\Lambda_{\rm hi} > \Lambda_{\rm med} > \Lambda_{\rm lo}$ according to Eq. (1). (The paths of searches using local minimization of the effective-energies is shown.) The lower panels roughly indicate the free-energy vs temperature trajectory diagrams that correspond to these potentials. (a) A "strong scaling" case where the search is "funneled" to the global minimum. (b) A "weak" scaling case where a single sequence of downhill searches does not find the global minimum. (From Shalloway, 1997.)

the free-energy global minimum will be found at the working temperature. Clearly, it will be important to determine the scaling properties of the energy landscapes of biopolymers of interest.

F. Overview

The fact that biopolymers fold on time-scales much shorter than those needed for a random search of conformation space suggests that they utilize specific, kinetically-favorable folding pathways to accelerate the process. Each pathway corresponds to a specific path down a macrostate trajectory diagram. The Packet Annealing Method is designed to mimic this efficient behavior by identifying and following the high-probability macrostates. The changes in macrostate position, size and number are traced using the characteristic packet equations. This is efficient because the effective-energy function is usually smooth within a single macrostate region. The approach has a number of unique features including:

1. *Physically appropriate:* it traces the free-energies of macrostates, not the energies of individual conformations. Thus, it accounts for not only the mean conformation, but also of the conformational fluctuations.

2. *Potential smoothing by the effective-energy:* the small fluctuations in the energy landscape are removed by the spatial averaging of Eq. (1). Therefore, minimization using $\widetilde{H}_K$ proceeds much more rapidly than minimization using V.

3. *Macrostate trajectory diagrams:* this novel description of the energy landscape uncovers its qualitative "hidden" structure and provides a "road-map" for organizing an efficient parallel search to the stable structure.

II. PROGRESS

A. Packet Annealing with Anisotropic Averaging Tensors

Our studies before the project had only used isotropic averaging tensors Λ_α^0 in which all fluctuations were assumed to be equal. However, the order-of-magnitude differences between the actual magnitudes of the fluctuations of different atom-pair distances in a macromolecule must be matched with anisotropic fluctuation tensors. We tested anisotropic averaging using the 6, 7 and 8 atom argon microclusters as test cases. Computational methods for iteratively solving the characteristic packet equations (3) and identifying the appearance of children at branch points (i.e., subsearch branching) were developed. This enabled us to compute macrostate trajectory diagrams for these systems, the first time that this has been done for non-trivial problems. For example, the 7-atom microcluster has 4 conformational isoforms (Fig. 3a) corresponding to local minima of the potential. While the 15 $(= 3 \times 7 - 6)$-dimensional potential $V(R)$ can not be visualized, the free-energy and mean-energy trajectory diagrams (Fig. 3b) reveal the hierarchical organization of the macrostates and associated isoforms. The critical temperatures (at the positions of the dotted lines) give the magnitudes of the effective energy barriers between these states. These studies, which demonstrated our ability to compute macrostates and trajectory diagrams for non-covalently linked systems, are discussed in more detail in Oresic and Shalloway (1994).

B. Packet Annealing in covalent systems

The effective-energy method was originally developed in Cartesian coordinates. but this is inefficient for biopolymer calculations because of the highly non-linear constraints imposed by the interatomic covalent bonds. Higher efficiency can be obtained using “internal coordinates”, a combination of bond-length, bond-angle and torsion-angle variables. A major goal (and accomplishment) of the project was to implement the method for covalently bonded biopolymers using internal coordinates.

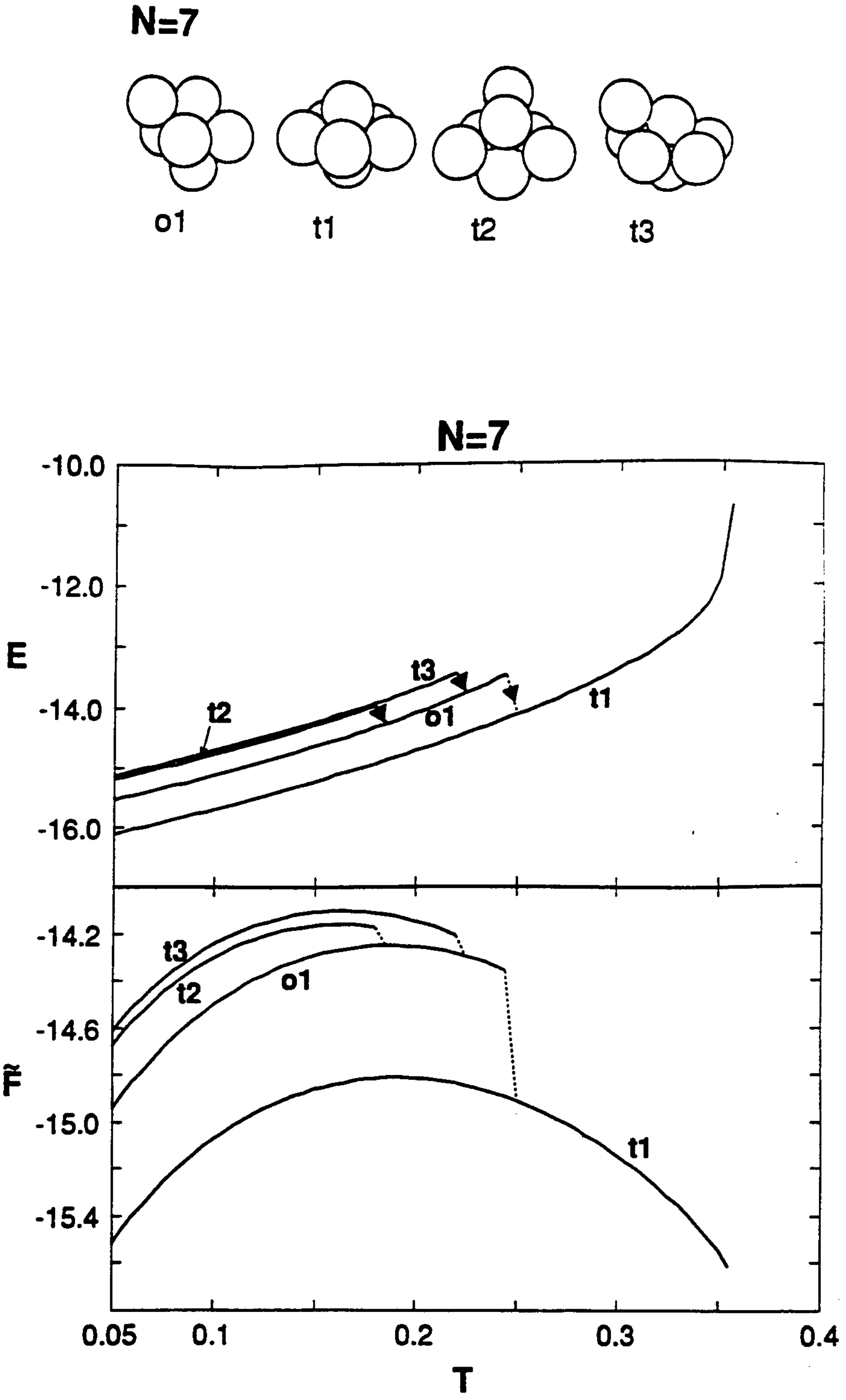

FIG. 3. Trajectory diagrams for 7-atom argon microclusters. (a) Isomers at $T = 0$. (b) Mean-energy (E) and free-energy (F) trajectory diagrams. The connections between trajectories (dotted lines and arrows) indicate transition temperatures where child macrostates merge with the parental states. The hierarchical organization of the landscape is apparent: isomers t_2 and t_3 are children of o_1 which itself is a child of t_1. (From Oresic and Shalloway, 1994.)

1. Multivariate wrapped-Gaussian distribution for biopolymers

We were surprised to find that a basic component of required mathematical theory, an accurate analog of the Cartesian coordinate Gaussian distribution for internal coordinates, had not yet been developed. (Most workers used the quasi-harmonic approximation even though this was very inaccurate for this application.) Thus, our first step was to develop the appropriate analog, the "multivariate wrapped-Gaussian distribution". A comparision of the performance of this method with the older quasi-harmonic method is shown in Fig. 4. This is a general-purpose tool which should find application in other biopolymer studies in addition to our own. See Church and Shalloway (1995, 1996).

2. Macrostate branching in distance-geometry variables

A major difficulty in analyzing biopolymers is that there are large differences between the effects of different torsion angles on conformation. Small changes in backbone angles near the center of a chain greatly affect conformation while changes in distal side-chains have little effect. This results in large matrix condition numbers that reduce efficiency. We showed that we could bypass this difficulty by describing the packets using pair-wise interatomic distance variables, i.e., distance-geometry coordinates. This representation works well because most metastable states can be distinguished using only a few distance variables so it is not necessary to consider all of the $O(N^2)$ variables (e.g., see Fig. 5). The description involves projecting the system probability density $p_\alpha(R)$ from torsion-angle variables Θ to the distance variables d:

$$p_\alpha^{ij}(d) = \int p_\alpha[R(\Theta)]\,|\nabla d_{ij}[R(\Theta)]|\,\delta(d - d_{ij}[R(\Theta)])\,\det^{\frac{1}{2}}[I(\Theta)]\,d\Theta \tag{5}$$

where

$$I(\Theta)_{np} \equiv \sum_i \frac{\partial R_i}{\partial \theta_n}\frac{\partial R_i}{\partial \theta_p}\,. \tag{6}$$

The division of a macrostate into child macrostates is readily identified algorithmically by

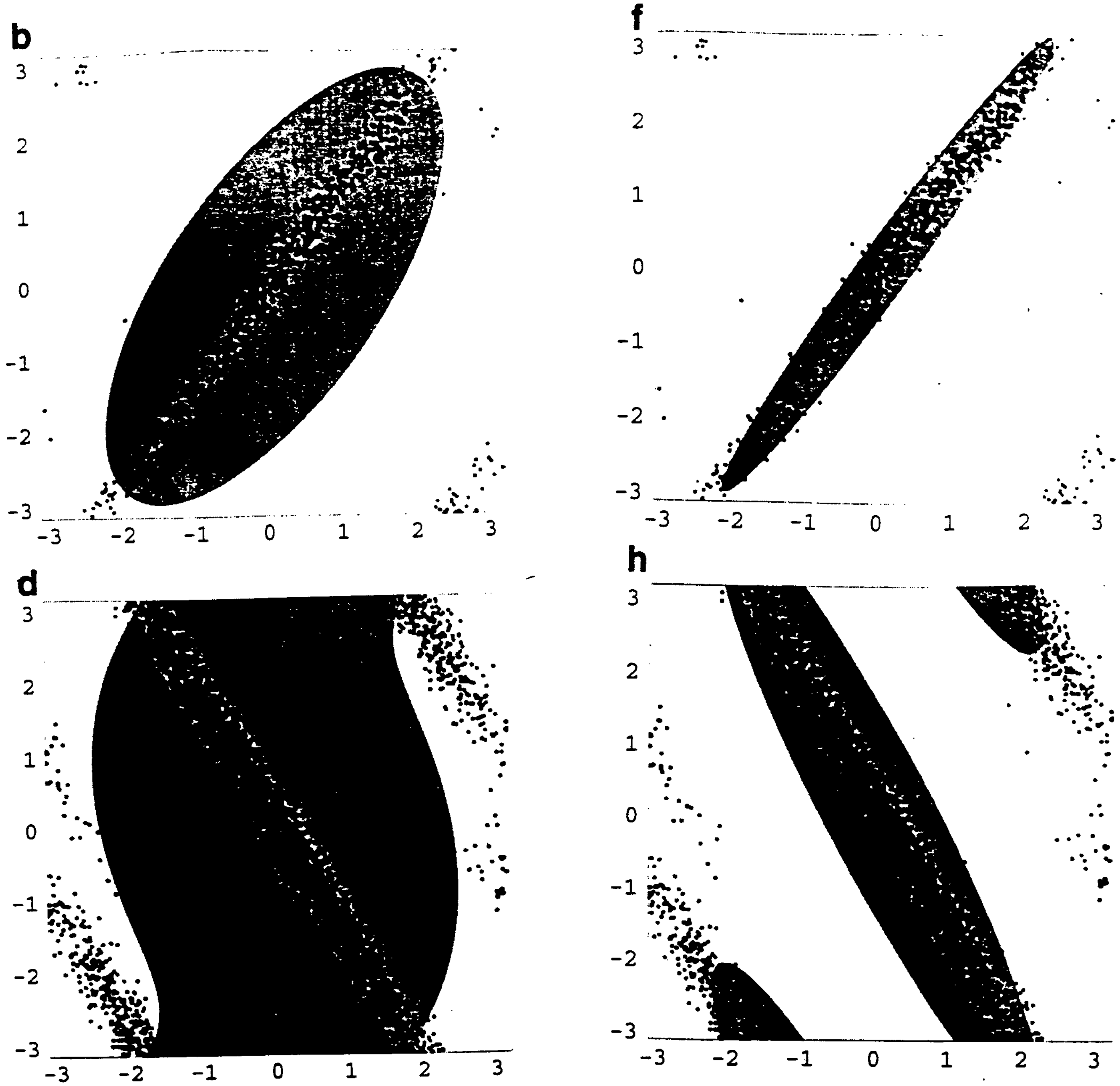

FIG. 4. Scatter plots of two-dimensional datasets having various amounts of correlation and fluctuation. The panels on the left (b and d) are shaded to show the e^{-2} regions generated by the angular quasiharmonic distribution; the panels on the right show the corresponding regions generated by the multivariate wrapped-Gaussian distribution. The significantly improved correspondence between the scatter plot and the shaded region reflects the higher accuracy of the wrapped-Gaussian distribution. (From Church and Shalloway, 1996.)

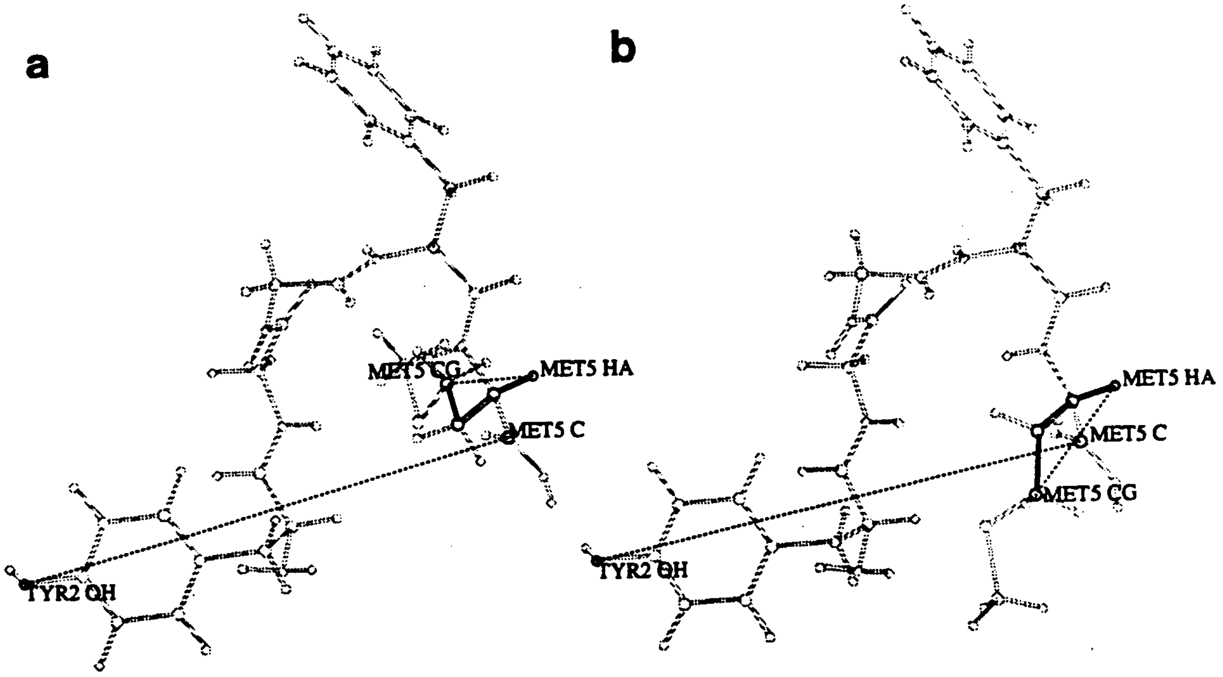

FIG. 5. Pairwise distances differentiate two conformations of Met-enkephalin. (From Church et al., 1996.)

inspecting the p_{α}^{ij} distributions (e.g., see Fig. 6). See Church et al. (1996a) and Church et al. (1998).

3. Trajectory diagram analysis of peptides

The methods described above enabled us to begin testing the method on the pentapeptide Met-enkephalin which has been used as a test-case for many theoretical studies. Our goal was to compute its trajectory diagram and to determine its scaling properties. The intrinsic parallelism of the effective-energy method provides an excellent opportunity for coarse-grained parallel computing. These studies were conducted using both the 16-processor Silicon Graphics Onyx and 512-processor IBM SP2 parallel computers available at the Cornell Theory Center. An algorithm was developed for automatically detecting the macrostate division points and assigning processors to the tracing of the low-lying trajectories. The Met-enkephalin probability and mean-energy trajectory diagrams are shown in Figs. 7 and 8. The probability trajectory diagram (Fig. 7) displays the Met-enkephalin macrostates having the highest probability (lowest free-energy) at each temperature as well as the trajectory which leads to the global energy minimum. It displays some features that will be common to all biopolymers: (1) At high temperature there is only one macrostate which fluctuates throughout all of conformation space (the peptide is denatured). (2) As temperature decreases, the total probability is distributed amongst an increasing number of macrostates. (3) At very low temperatures, all of the probability becomes concentrated in the macrostate containing the global energy minimum. It is important to note that the trajectory that leads to the global minimum at $300°K$ ($\log T \approx -0.22$ in the units used in Fig. 7) is not the highest probability trajectory at intermediate temperatures (e.g., $-0.1 < \log T < 0.1$), indicating that Met-enkephalin has weak-scaling in free-energy. However, an important discovery was that it has strong scaling in mean-energy: as seen in Fig. 8 the global minimum trajectory remains near the bottom of the mean-energy trajectory diagram for all temperatures. Furthermore, there is a gap between the global minimum trajectory and the other

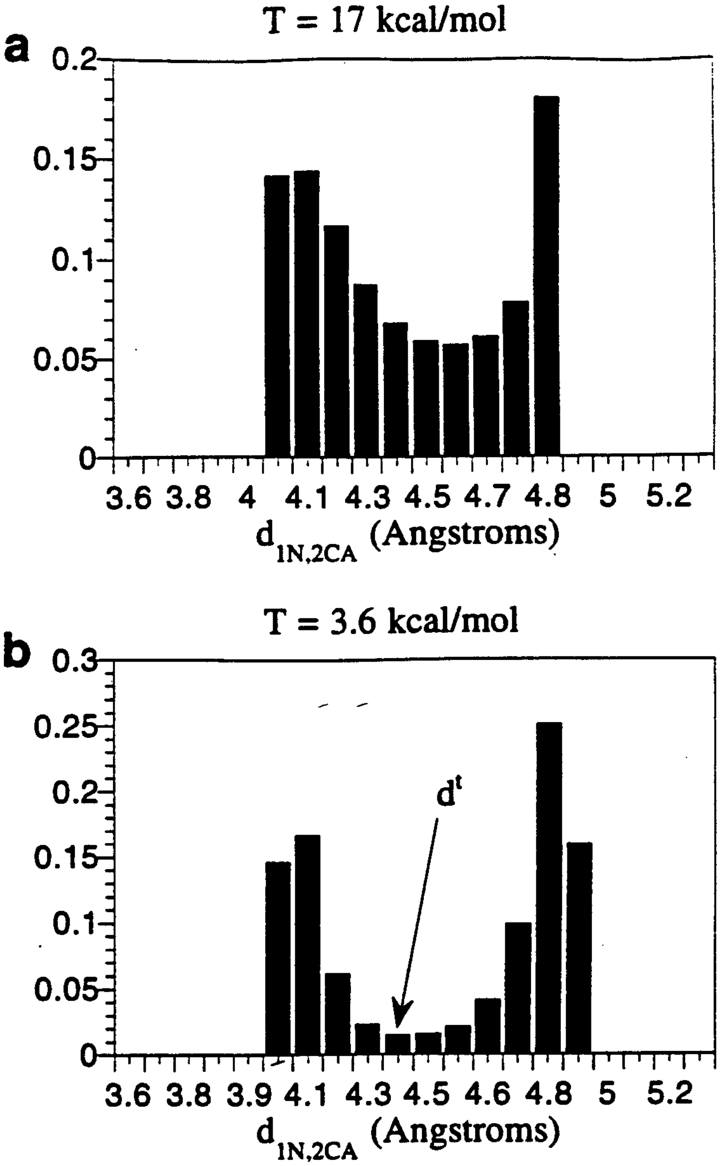

FIG. 6. Projected probability distribution $p^{1N,2CA}$ at high temperature (a) and at low temperature (b). At the low temperature the distribution satisfies the criteria for bifurcation into two macrostates: one with $d_{1N,2CA} < d^t$ and one with $d_{1N,2CA} > d^t$. (From Church et al., 1996.)

trajectories over a significant temperature range. This means that one can easily find the Met-enkephalin global minimum by just tracking a small number of macrostate trajectories having the low mean-energy.

We have performed similar studies with other peptides to test the generality of this phenomenon. Interestingly, the mean-energy trajectory diagram of Leu-enkephalin, which differs from Met-enkephalin by a single amino acid substitution, displays only weak scaling and does not have the energy gap. This implies that global minimization of Leu-enkephalin should be more difficult than minimization of Met-enkephalin. This prediction has been verified by comparing the simulated annealing of the two molecules. See Church et al. (1996) and Church et al. (1998).

Most recently, we have developed code to implement the modified image electrostatics (MIMEL)[7] approach to empirical solvation and incorporate it into our model.

C. Dynamical basis for effective-energy global optimization

We showed how the characteristic packet equations emerge from stochastic analysis of macromolecular motions using the Smoluchowski (Fokker-Planck) equation. This leads to a time-dependent description of the system conformational probability distribution that is analogous to the wave-function description of the Schrodinger equation. These studies showed how the Packet Annealing Method is related to the macromolecular dynamics and led to a new, efficient variational method for computing transition rates between macromolecular conformational states (see below). This provides an important extension to the global minimization problem by allowing us to compute the rates of folding and conformational change. See Shalloway (1996).

1. Variational calculation of conformational transition rates

At the present time conformational reaction rates are generally approximated by transition state and reactive flux methods (Ref. 8, for review). These methods require the

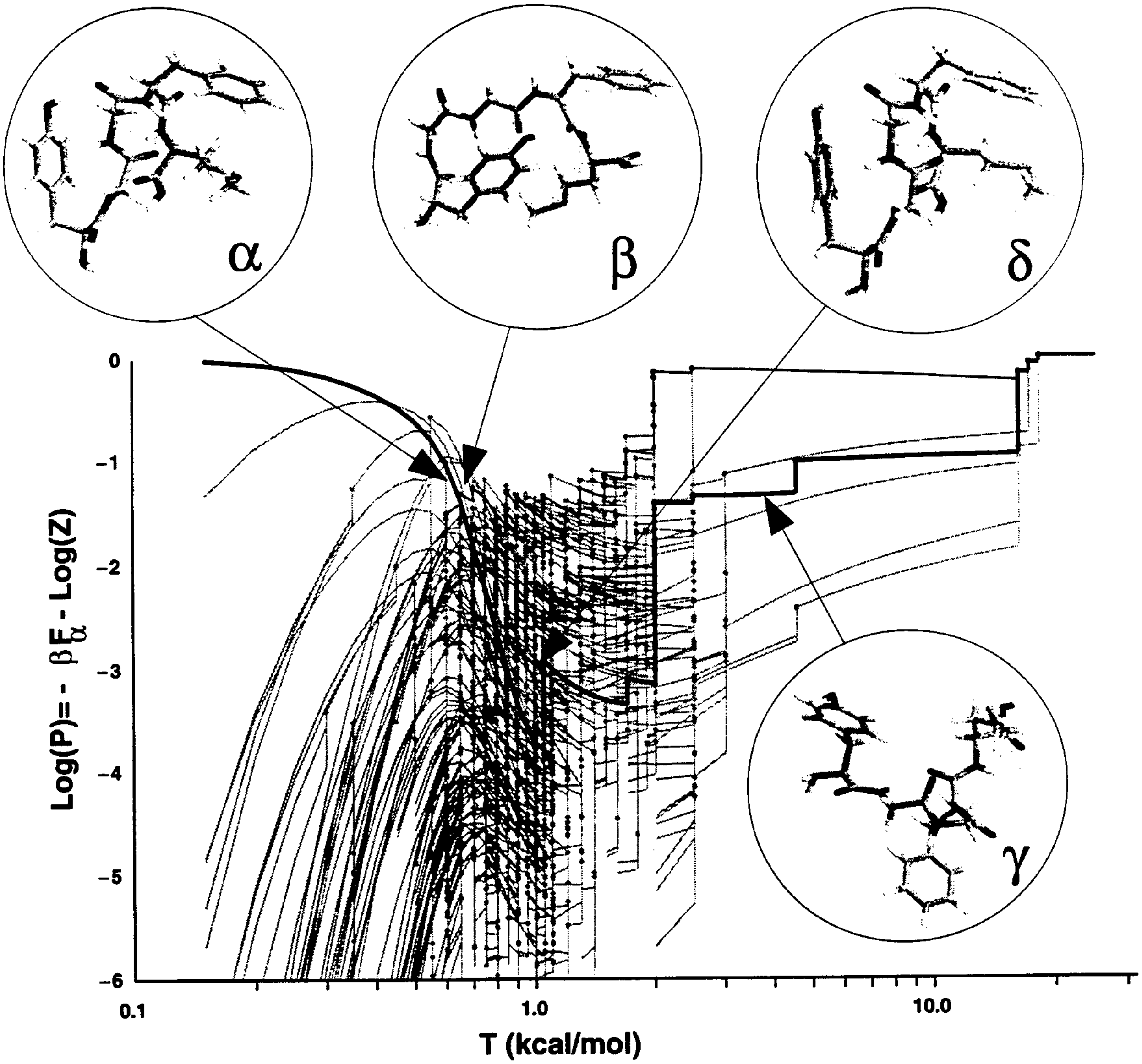

FIG. 7. Peptide trajectory diagrams. The high-probability trajectories in the probability trajectory diagram of Met-enkephalin. The trajectory which leads to the global minimum (marked with an ×) is also shown. The fact that its probability goes down to $\sim 10^{-4}$ at $\log T \sim 0.1$ implies that the scaling is very weak in probability (or, equivalently, free energy). (From Church and Shalloway, 1998.)

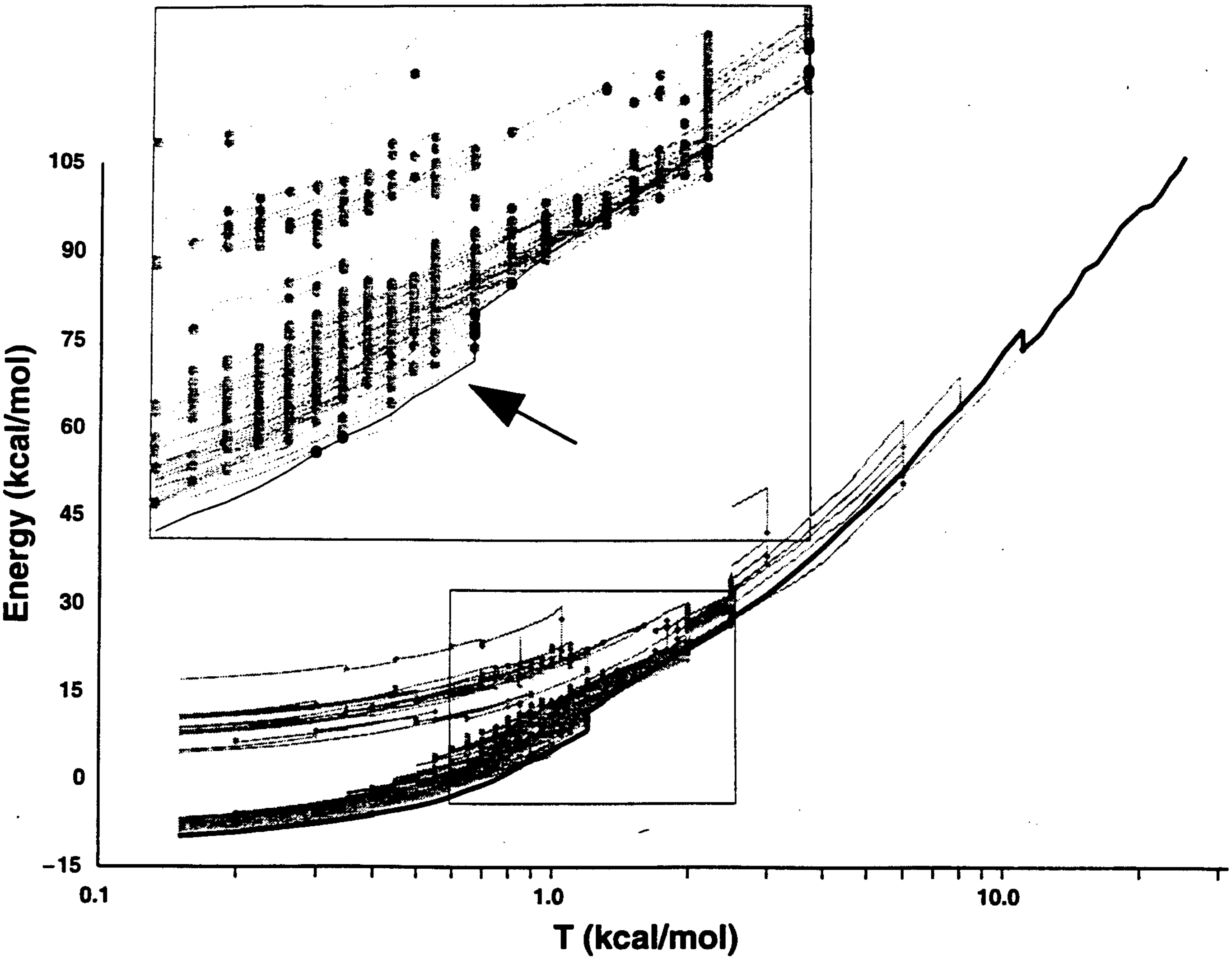

FIG. 8. Low mean-energy trajectories for Met-enkephalin. The trajectory which leads to the global minimum has the lowest, or close to lowest, mean-energy at all temperatures—an example of strong scaling. Note also the mean-energy gap between this and the other trajectories in the complicated $\log T \sim -0.2$ region. (From Church and Shalloway, 1998.)

specification of one-dimensional "reaction coordinates" that can describe the progression of the system from one conformational macrostate to another. However, there is no general method for identifying appropriate reaction coordinates, and it is very difficult, if not impossible, to find them for complex multidimensional systems like macromolecules. Even when they can be found, rate computations are exceedingly expensive and often inaccurate.

The Smoluchowski formulation mentioned above leads to a reaction path-independent method for computing transition rates that can avoid these difficulties. Based on the quantum-mechanical analogy, we have shown that transition rates can be efficiently determined by using the Rayleigh-Ritz variational principle to compute the eigenvalues of the first excited states of the Smoluchowski "hamiltonian".

We have developed and tested this variational method using model potentials and the argon microcluster system as test cases. Computational methods for iteratively solving the variational equations were developed and tested, and showed that the method was about two orders-of-magnitude more efficient than Brownian dynamics for equal accuracy. (Ulitsky and Shalloway, 1998). As part of this project we have developed a new "contangency" method for finding saddle points of effective energy landscapes (Ulitsky and Shalloway, 1997) which can be used by other transition rate computation methods as well.

D. Relationship between the Packet Annealing, Diffusion Equation, and Gaussian Density Annealing methods

Our effective-energy method and the competing Diffusion Equation[9,10] and Adiabatic Gaussian Density Annealing[11,12] methods all use Gaussian convolutions to smooth the macromolecular potential. However, the relationship bvetween these methods has not previously been understood. We have now shown that they are hierarchically related: the Diffusion Equation Method is a special case of the Gaussian Density Annealing method (restriction to isotropic averaging), and the Gaussian Density Annealing method is in turn a special case of the Packet Annealing Method (restriction to anisotropies along fixed axes, single packets, and high temperature). Thus, the Packet Annealing Method provides a general formulation for this entire class of models. See Shalloway (1997).

E. Additional progress

1. Spatial interpolation integrators for molecular dynamics simulation

Molecular dynamics simulations are much less efficient than the effective-energy techniques discussed above but are useful for studying the details of fast conformational transitions. They are most commonly performed using the Verlet algorithm to integrate Newton's equation. However, this is a general-purpose integrating algorithm which does not exploit an important special property of Newton's equation for biopolymers—that the force is the gradient of a scalar potential. We have shown that a new class of *spatial interpolation algorithms*, which do exploit this property, can enhance efficiency by factors of 4–5. See Gueron and Shalloway (1996).

2. Benchmark for molecular dynamics simulations

In developing the spatial interpolation method we recognized that the existing methods for evaluating the accuracy of molecular dynamics simulations were inadequate. While energy-conservation is often used, this is a crude measure that does not accurately reflect performance. To solve this problem we developed a new method which uses the "residual force" $\nabla V(R)+m\,d^2R/dt^2$ (where V is the is the potential and R is the computed trajectory) as a measure of accuracy. A software package supporting this benchmark is being prepared for general release. See Gans et al. 1998.

3. Correlation between codon usage and protein secondary structure

Biopolymer production in most cases involves expression of modified genes in heterologous bacterial or plant expression systems. Genes are constructed assuming that it does not matter which of the synonymous codons that encode a specific amino acid are used. While evidence indicates that this assumption is usually true, it may not always be true, and specific synonymous codon choices might be important for directing protein folding. This (controversial) hypothesis might explain the inability of some genes to be expressed in heterologous sources. We statistically analyzed protein and nucleic acid databases to test this hypothesis and found strong evidence for correlation between some synonymous codons and protein structures. See Oresic et al., 1998.

III. EXECUTIVE SUMMARY

A. Publications

1. Pardalos, P., D. Shalloway, and G. Xue (1994) Optimization methods for computing global minima of nonconvex potential energy functions. *J. Global Optimization* 4:117-133.

2. Coleman, T., D. Shalloway, and Z. Wu (1994) A parallel build-up algorithm for global energy minimizations of molecular clusters using effective energy simulated annealing. *J. Global Optimization* 4:171–186.

3. Oresic, M., and D. Shalloway (1994) Hierarchical characterization of energy landscapes using Gaussian packet states. *J. Chem. Phys.* 101:9844–9857.

4. Church, B., and D. Shalloway (1995) Characterizing large correlated fluctuations of peptide conformations in torsion angle space. *Polymer Prep.* 36:636-637.

5. Church, B.W., M. Oresic and D. Shalloway (1996) Tracking metastable states to free-energy global minima, in *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, v. 23, P. Pardalos, D. Shalloway and G. Xue, eds. (Amer. Math. Soc., Providence, R.I.), pp. 41–64.

6. Church, B., and D. Shalloway (1996) Characterizing large correlated fluctuations of macromolecular conformations in torsion-angle space using the multivariate wrapped-Gaussian distribution. *Polymer* 37:1805-1813.

7. Gueron, S., and D. Shalloway (1996) Spatial interpolation methods for integrating Newton's equation. *J. Comp. Phys.* 129:87–100.

8. Shalloway, D. (1996) Macrostates of classical stochastic systems. *J. Chem. Phys.* 105:9986–10007.

9. Ulitsky, A., and D. Shalloway (1997) Finding transition states using contangency curves. J. Chem Phys. 106:10099–10104.

10. Shalloway, D. (1997) Variable-scale coarse-graining in macromolecular global optimization. in *Large Scale Optimization with Applications to Inverse Problems, Optimal Control and Design, and Molecular and Structural Optimization*, L. T. Biegler, T. F. Coleman, A. R. Conn, and F. Santosa, eds. (Springer, New York) pp. 135–161.

11. Church, B. W. A. Ulitsky, and D. Shalloway (1998) Macrostate dissection of thermodynamic Monte Carlo integrals. Adv. Chem. Phys. (in press).

12. Shalloway, D. Packet annealing. In *Encyclopedia of Global Optimization*, C. A. Floudas and P. M. Pardalos, eds. (Kluwer Academic, Norwell, MA) (in press).

13. Ulitsky, A., and D. Shalloway (1998) Variational calculation of macrostate transition rates. J. Chem. Phys. (in press).

14. Oresic, M. and D. Shalloway (1998) Specific correlations between relative synonymous codon usage and protein secondary structure. (submitted)

15. Gans, J., J. Chan, and D. Shalloway (1998) Residual force as a benchmark for the accuracy of molecular dynamics simulations. (in preparation)

B. Meeting presentations

1. Oresic, M., and D. Shalloway (1994) Hirerachical characterization of energy landscapes using Gaussian packet states. Symposium on Computation in Biophysical Chemistry, Cornell Theory Center..

2. Church, B. W., and D. Shalloway (1995) Characterizing large correlated fluctuations of peptide conformations in torsion angle space. American Chemical Society Meeting.

3. Church, B. W., M. Oresic and D. Shalloway (1995) Tracking metastable states to free-energy global minima. DIMACS Miniworkshop on Global Minimization of Nonconvex Energy Functions.

4. Shalloway, D., B. W. Church and M. Oresic (1995) Computing the hierarchical structure of peptide folding pathways. Ninth Meeting of Groups Studying the Structure of AIDS-Related Systems and Their Application to Targeted Drug Design.

5. Shalloway, D. (1995) Hierarchical analysis of energy landscapes for biopolymer structure prediction. AFOSR Meeting on Optimization of Molecular Structures.

6. Shalloway, D. (1995) Variable-scale coarse-graining in macromolecular global optimization. Program on Large Scale Optimization with applications to Inverse Problems, Optimal Control and Design, and Molecular and Structural Optimization, Institute for Mathematics and Its Applications (U. Minn.).

7. Shalloway, D. (1995) Computing protein folding pathways. Symposium on Protein Structure and Folding, Cornell Theory Center.

8. Church, B. W., J. Gans, M. Oresic, A. Ulitsky and D. Shalloway (1996) Hierarchical analysis and effective energy methods for protein free-energy global minimization. Fifth SIAM Conference on Optimization.

9. Shalloway, D., and M. Oresic (1996) Correlation between synonymous codon usage and protein secondary structure. Eastern Great Lakes Molecular Evolution Meeting.

10. Church, B. W. (1996) Computing protein conformational hierarchies. American Chemical Society Meeting.

11. Shalloway, D. (1997) Hierarchical mapping of macromolecular conformational landscapes using macrostate trajectory diagrams. Second International Symposium "Algorithms for Macromolecular Modelling," Konrad-Zuse-Zentrum, Berlin.

12. Shalloway, D. (1997) Hierarchical structure of macromolecular energy landscapes. Seventy-eighth Statistical Mechanics Conference.

In addition, six seminars on this project were presented at universities in the US, Europe and Korea.

C. Scientific personnel involved

1. Prof. David Shalloway

2. Prof. Tom Coleman

3. Dr. Bruce Church (postdoc)

4. Dr. Shay Gueron (postdoc)

5. Dr. Alex Ulitsky (postdoc)

6. Dr. Zhijun Wu (postdoc)

7. Jason Gans (graduate student)

8. Matej Oresic (graduate student)

9. Linda Rosenband (graduate student)

REFERENCES

[1] S. Kirkpatrick, J. C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).

[2] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Addison–Wesley, Reading, MA, 1969).

[3] R. Elber and M. Karplus, Science **235**, 318 (1987).

[4] F. H. Stillinger and T. A. Weber, Phys. Rev. A **25**, 978 (1982).

[5] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).

[6] C. Levinthal, J. Chim. Phys. **65**, 44 (1968).

[7] R. Abagyan and M. Totrov, J. Mol. Biol. **235**, 983 (1994).

[8] P. Hänggi, P. Talkner, and M. Borkovec, Rev. Mod. Phys. **62**, 251 (1990).

[9] L. Piela, J. Kostrowicki, and H. Scheraga, J. Phys. Chem. **93**, 3339 (1989).

[10] J. Kostrowicki and H. A. Scheraga, in *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, Vol. 23 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, edited by P. Pardalos, D. Shalloway, and G. Xue (American Mathematical Society, Providence, RI, 1996), pp. 123–132.

[11] J. Ma and J. E. Straub, J. Chem. Phys. **101**, 533 (1994).

[12] P. Amara, J. Ma, and J. E. Straub, in *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, Vol. 23 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, edited by P. Pardalos, D. Shalloway, and G. Xue (American Mathematical Society, Providence, RI, 1996), pp. 1–13.