Deconstructing the energy landscape: Constraint-based algorithms for folding heteropolymers

Veit Elser* and Ivan Rankenburg
Department of Physics, Cornell University, Ithaca, New York 14853, USA
(Received 21 July 2005; published 6 February 2006)

We apply the computational methodology of phase retrieval to the problem of folding heteropolymers. The ground state fold of the polymer is defined by the intersection of two sets in the configuration space of its constituent monomers: a geometrical chain constraint and a threshold constraint on the contact energy. A dynamical system is then defined in terms of the projections to these constraint sets, such that its fixed points solve the set intersection problem. We present results for two off-lattice hydrophobic-polar models: one with only rotameric degrees of freedom, and one proposed by Stillinger *et al.* [Phys. Rev. E **48**, 1469 (1993)] with flexible bond angles. Our phase retrieval inspired algorithm is competitive with more established algorithms and even finds lower-energy folds for one of the longer polymer chains.

DOI: 10.1103/PhysRevE.73.026702 PACS number(s): 05.10.−a, 87.15.Aa, 87.15.Cc

I. INTRODUCTION

A favorite metaphor in the field of nonlinear optimization, and computational protein folding in particular, is the *energy landscape*. Energy landscapes have been compared to funnels [1] and golf courses [2], and are generally held responsible for all the behavior observed in nature, as well as the challenges faced by simulators. Kinetics simulations are, by their very nature, tied to the topography of the energy landscape and cannot avoid scaling its barriers and languishing in its manifold minima. The outlook for native fold discovery, however, is more optimistic. As we show below, for this problem there are options that escape the confines of the energy landscape and yield significant computational dividends.

Most native fold search strategies are conservative in at least two respects. First, the search is carried out in the same space accessed by the physical degrees of freedom of the protein. Second, the search in this space is carried out quasilocally, in the sense that every conformation examined is derived from a previously considered conformation by a local modification. There are alternatives to these general guidelines that have proven effective in other fields. For inspiration we turn to the classic problem of *phase retrieval*.

The naive search space in phase retrieval is superficially equivalent to the space of rotamer configurations, each unknown phase angle ϕ corresponding to a dihedral angle on the protein backbone. An important application of phase retrieval is the reconstruction of the electron density in a crystal, given its Fourier amplitudes $F_{\mathbf{q}}$:

$$\rho(\mathbf{r}) = \sum_{\mathbf{q}} F_{\mathbf{q}} \cos(\mathbf{q}\cdot\mathbf{r} + \phi_{\mathbf{q}}). \tag{1}$$

The task of the algorithm is to find values for the phases $\phi_{\mathbf{q}}$ such that the resulting density (1) satisfies certain general characteristics (e.g., positivity, atomicity) or constraints. To illustrate the idea, we consider a very simple situation where the given amplitudes $F_{\mathbf{q}}$ are derived from a density known to take only two values, say $\rho = \pm 1$. To implement the binary valued density constraint we could try minimizing a penalty function of the form

$$V = \sum_{\mathbf{r}} [\rho(\mathbf{r})^2 - 1]^2, \tag{2}$$

where the positions $\mathbf{r}$ fall on a grid determined by the range over which the Fourier vectors $\mathbf{q}$ are sampled. This expression for V, an explicit function of the phase variables $\phi_{\mathbf{q}}$, is a possible energy landscape for the phase retrieval problem. The correct phases are identified by discovering a point on the landscape where the energy realizes the minimum value $V=0$.

Practical phase retrieval algorithms do not minimize an objective function as sketched above [3–5]. The most successful algorithms do not navigate the barriers and false minima of an energy landscape. Typically, the search performed by these algorithms is carried out in a much larger space (than the space of "rotamers") and the steps executed are global in character. The example above serves to illustrate the key elements of the search dynamics, called *projections*. There are two projections, both of which act on a density that has been freed of all constraints. In particular, one no longer insists that ρ has the given Fourier amplitudes, that is, the form (1) parametrized by phase angles. Instead, one uses the device of a projection P_A, which takes an arbitrary input density ρ and returns a minimal modification of ρ where the given Fourier amplitudes have been restored. This can be computed efficiently, by first transforming ρ to Fourier space, making the necessary modification there, and then transforming back. The term "projection" is derived from the minimality condition, and in the case of P_A corresponds (in Fourier space) to mapping each complex Fourier coefficient to the nearest point on a circle whose radius is given by the corresponding amplitude $F_{\mathbf{q}}$. The binary constraint on the density values is implemented by another projection P_B, where minimality of the change calls for all positive values to be replaced by 1, negative values by −1. Each of the two projections accomplishes something global, in effect solving half of the problem to completion. The spectrum of modern phase retrieval algorithms arises from both the variety of the

*Electronic address: ve10@cornell.edu

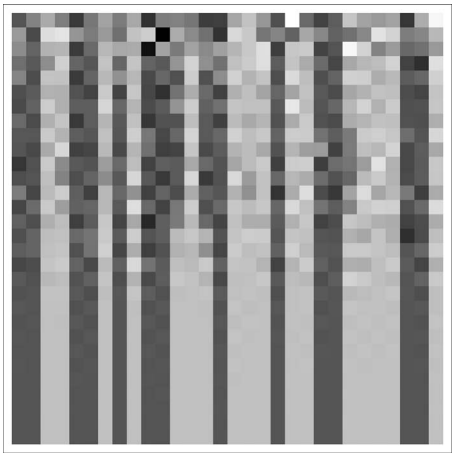

FIG. 1. The difference map solution of a phase retrieval problem resembles a deterministic cellular automaton. The cells in each horizontal row represent one iterate of the one-dimensional, reconstructed density. Starting from any initial density (top), iteration of the map eventually arrives at the fixed point solution (bottom).

kinds of projections used, as well as variations in how they are combined [6].

Figure 1 shows successive iterates of a particular combination of projections, called the *difference map* [6], on the phase retrieval problem for a binary valued density. The dynamics is deterministic and the discovery of the solution corresponds to the arrival at a fixed point of the map. Although the number of iterations required by the algorithm depends on the initial density, this number is always much less than the size of an exhaustive search.

We show below that the projection technique can be applied to the protein native fold search problem, and that for simple off-lattice heteropolymer models the results are encouraging. After a brief review of the difference map scheme for combining projections, we examine in detail the two projections that apply to the native fold search. We present results for two hydrophobic-polar (HP) models, one with only dihedral degrees of freedom (rotamer model), and a model proposed by Stillinger *et al.* [7] with variable bond angles (flexible chain model). For the longer chains the projection based algorithm was able to find lower energies than published results [8,9] obtained by methods that explore the energy landscape.

II. THEORY AND MODELS

A. Difference map algorithm

The search space is in general a high-dimensional Euclidean space E. Polymer conformations, for example, are embedded by associating three Cartesian coordinates of E with the position of each monomer in the chain. The goal of the algorithm is to discover one element $x \in A \cap B$, where A and B are subsets of E, usually having the character of constraints. In polymer applications, for example, set A might represent all monomer configurations that satisfy the chain constraints (bond lengths, etc.). The constraint sets A and B are assumed to be simple enough that the two projections to these sets, P_A and P_B, can be computed efficiently. For example, to compute $y=P_A(x)$, we need to find an element $y \in A$ that minimizes the distance $\|y-x\|$. In difference map applications one may relax the condition that $y \in A$ realizes the true minimum of $\|y-x\|$, although this is usually easy to achieve when y is near enough to x that the constraint can be linearized. In general, the performance of the algorithm is improved by the distance minimizing quality of the projections.

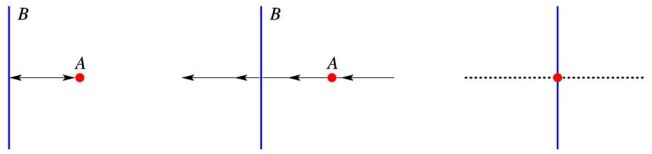

FIG. 2. (Color online) Comparison of alternating projections (left) and difference map iterations (center) in the case of two constraint sets, a point (red) and a line (blue), that do not intersect. The alternating map $P_A(P_B(x))$ stagnates on set A; iterates of $D(x)$ move uniformly along the axis of nearest separation between A and B. When A and B intersect (right), every point in the space locally orthogonal to both constraints is a fixed point of $D(x)$.

When the projections are combined in alternating fashion, $x \mapsto P_A(P_B(x))$, problems arise when there is a local minimum in the separation of the constraint sets. As shown in Fig. 2, this map can then have a fixed point $x^*=P_A(P_B(x^*))$ that lies in A but not B. The difference map is a more elaborate combination of projections given by [6]

$$x \mapsto D(x) = x + \beta\Delta(x), \tag{3}$$

$$\Delta(x) = P_A(f_B(x)) - P_B(f_A(x)), \tag{4}$$

where

$$f_A(x) = P_A(x) - \beta^{-1}[P_A(x) - x], \tag{5}$$

$$f_B(x) = P_B(x) + \beta^{-1}[P_B(x) - x], \tag{6}$$

and $\beta \neq 0$ is a dimensionless parameter. At a fixed point $x^*=D(x^*)$, we have $\Delta(x^*)=0$ and

$$P_A(f_B(x^*)) = P_B(f_A(x^*)) = x_{\mathrm{sol}}. \tag{7}$$

This shows that $x_{\mathrm{sol}} \in A \cap B$, since x_{sol} is in the range of both projections. The more straightforward definition $\Delta(x) = P_A(x) - P_B(x)$, which leads to the same conclusion, is not useful because the fixed points x^* of D are then unstable. The maps f_A and f_B are tuned to maximize the attraction of the difference map's fixed points [10]. When confronted with a near intersection of sets A and B, iterates of the difference map move at a uniform rate along the axis of nearest separation, as shown in Fig. 2. The step size in the latter situation decreases in proportion to the distance between A and B, and the flow degenerates into a space of fixed points when A and B intersect.

Studies of hard optimization problems, such as phase retrieval, point to the following sequence of events in the difference map solution process. Starting from an arbitrary initial point $x_0 \in E$, the iterates very quickly converge on a much smaller subset, a quasiattractor Q. The dynamics on Q

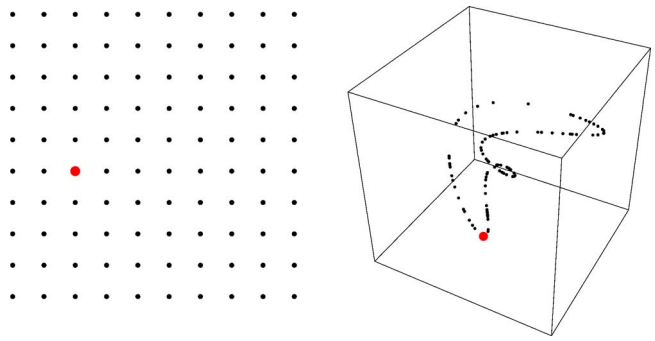

FIG. 3. (Color online) Cartoon of the spaces sampled by optimization algorithms: rotamer space for the case of two dihedral angles (left), and the difference map quasiattractor (right). The dimension of the quasiattractor is smaller than that of the corresponding rotamer space, even though it is embedded in a higher-dimensional Euclidean space. The large (red) point represents the solution.

is chaotic, and Q would be a true (chaotic) attractor in an ill-posed problem instance, when $A \cap B$ is empty. Since the two projections are in fact very insensitive to the existence of a solution, it follows that the dynamics in a well-posed instance is similar, only differing when the iterate arrives at the attractive basin of a fixed point and the algorithm terminates. A cartoon comparison of exhaustive rotamer search and difference map search is given in Fig. 3.

B. Heteropolymer models

We consider two off-lattice heteropolymer models, with monomer-monomer interaction of the Lennard-Jones form:

$$E_{\mathrm{LJ}} = 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N}\left(\frac{1}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}\right). \tag{8}$$

N is the number of monomers, $\mathbf{r}_{ij}$ is the vector separation of monomers i and j with magnitude $|\mathbf{r}_{ij}| = r_{ij}$, and $C_{ij} = C_{ji}$ are constants that depend on the hydrophobic (H) and polar (P) character of the monomers. For the *flexible chain* model proposed by Stillinger *et al.* [7],

$$C_{\mathrm{HH}} = 1, \quad C_{\mathrm{HP}} = -\frac{1}{2}, \quad C_{\mathrm{PP}} = \frac{1}{2}. \tag{9}$$

Another model we study, the rotamer model, has

$$C_{\mathrm{HH}} = 1, \quad C_{\mathrm{HP}} = C_{\mathrm{PP}} = \frac{1}{2}. \tag{10}$$

The main difference between the flexible chain and rotamer models is the nature of the constraints on the polymer chain. In the flexible chain model only the bond length is fixed, $r_{ii+1} = 1$; in the rotamer model the bond angles are fixed as well: $\mathbf{r}_{i-1i} \cdot \mathbf{r}_{ii+1} = \cos\alpha$. Since the latter constraint fixes the distances r_{ii+2}, these terms are excluded from the sum in (8) for the rotamer model. The flexible chain model adds a bond angle energy favoring linear conformations:

$$E_{\mathrm{chain}} = \frac{1}{4}\sum_{i=2}^{N-1}(1 - \mathbf{r}_{i-1i} \cdot \mathbf{r}_{ii+1}). \tag{11}$$

C. Constraint projections

Protein conformations are subject to two, typically antagonistic, constraints. In order to function, proteins adopt a compact shape with stability and functionality conferred by the three dimensional packing of its constituent amino acid residues. In order for the protein to be synthesized, however, the arrangement of the residues must also correspond to a possible conformation of a polypeptide chain. Either of these constraints would be much easier to satisfy if the other could be neglected, and there would then be a multitude of solutions. The difficulty in finding the native fold, from this perspective, is finding a configuration of residues that satisfies both constraints. We discuss later how this point of view provides a basis for understanding the uniqueness of the native fold.

The application of the difference map algorithm to the model proteins described above involves three things: specifying the embedding, defining the constraint sets, and computing projections to the constraint sets. We embed both models in a Euclidean space E of dimension $3N$ in the standard way: three Cartesian coordinates for each monomer position. The constraint sets A and B correspond to the chain constraints and packing constraints, respectively.

Set A in the flexible chain model is the set of all monomer configurations in E with $r_{ii+1} = 1$, while in the rotamer model we impose the additional constraint $\mathbf{r}_{i-1i} \cdot \mathbf{r}_{ii+1} = \cos\alpha$ (for a given α). The projection to A, or P_A, is computed with the aid of a penalty function V_{chain}. For the rotamer model we used

$$V_{\mathrm{chain}} = \sum_{i=1}^{N-1}(r_{ii+1} - 1)^2 + \sum_{i=2}^{N-1}(\mathbf{r}_{i-1i} \cdot \mathbf{r}_{ii+1} - \cos\alpha)^2. \tag{12}$$

The flexible chain model used only the first term in (12). To compute $P_A(x)$, given some input monomer configuration $x \in E$, we use gradient descent minimization of V_{chain}, terminated when the step size falls below a given threshold. The algorithm records the success of the projection by testing whether V_{chain} is within a small tolerance value of zero. In the experiments reported below, the success rate for P_A was 100%.

The packing constraint set B in the rotamer model is simply the set of monomer configurations $x \in E$ satisfying $E_{\mathrm{LJ}}(x) < E_0$, where E_0 specifies the energy depth of the search. For the constraint satisfaction problem to have a feasible point, or the difference map to have a fixed point, E_0 must be greater than the ground state energy of the polymer. We again compute the corresponding projection, P_B, using gradient descent, but now with the function E_{LJ}. The termination criterion is also different, since we are only interested in crossing the $E_{\mathrm{LJ}}(x) = E_0$ contour, rather than finding a local minimum. After crossing the target contour, we use Newton iterations to converge on the contour, as this brings us slightly closer to the input, while still satisfying the constraint. In the event that the input x already satisfies $E_{\mathrm{LJ}}(x) < E_0$, the same x is returned as the output of the projection. Crossing of the E_0 contour is used as the criterion for a successful computation of P_B. Clearly the success rate depends on E_0. In our experiments the success rate for P_B was

essentially 100%, since the target energy E_0 is always such that finding a feasible point of $E_{\mathrm{LJ}}(x) < E_0$ is easy. This is because the target energies of relevance, those that apply in the dual constraint problem, are always significantly above the minimum energy of the pure packing problem (no chain constraint). Because the inputs to projections generated by the difference map scheme can fall within regions where E_{LJ} diverges sharply, we modified the Lennard-Jones potential to have the form $a - br_{ij}^2$ for separations $r_{ij} < 0.9$, with a and b chosen to make E_{LJ} and its first derivatives continuous. All the folds discovered by the algorithm have $r_{ij} > 0.9$ for all monomer pairs i and j. A similar soft-atom modification of the Lennard-Jones function was used by Levitt [11].

Figure 4 shows the action of the chain constraint projection, P_A, on a configuration of monomers in the rotamer model with $\cos\alpha = 0.5$. The H-P sequence is known only to P_A; in this example it is periodic with a three-element motif: $(\mathrm{HPP})_8$. The packing constraint P_B is blind to the sequence ordering of monomers.

The formulation of the packing constraint set, and the computation of its projection, was somewhat different in the flexible chain model. This example illustrates both the pitfalls in the naive application of the difference map algorithm, as well as its flexibility. The chain energy (11) would seem to have its natural place in defining the chain constraint A. However, this would entail having to specify another adjustable energy parameter in addition to the packing energy E_0. The other option, of combining E_{chain} with E_{LJ} (thereby modifying set B), would be a mistake because the former has a very long-range character, in contrast with the latter, and the projection would almost always be blind to the possibility of favorable monomer contacts. Our solution was to combine a modified form of E_{chain} with E_{LJ}:

$$E'_{\mathrm{chain}} = \frac{1}{4}\sum_{i=2}^{N-1} (1 - \mathbf{r}_{i-1i} \cdot \mathbf{r}_{ii+1}) w(r_{i-1i}) w(r_{ii+1}), \qquad (13)$$

where

$$w(r) = \begin{cases} 1, & \text{if } r \leqslant 1, \\ 1 - (1/r^2 - 1)^2, & \text{if } r > 1. \end{cases} \qquad (14)$$

A modification of this kind is valid, since any solution $x \in A \cap B$ satisfies the chain constraint, and E'_{chain} reduces to E_{chain}.

Gradient descent to a constraint set specified by the contours of a function, is only distance minimizing when the constraint function is linear. We considered the possibility, when seeking a nearby point on a contour, say $V(x) = V_0$, that it may be advantageous to perform gradient descent on a "guiding function," say $G(x)$. The descent would still be terminated at the contour of the original function; the role of the guiding function is only to minimize the length of the path to the contour. In the rotamer model we obtained good quality projections without the use of guiding functions. In the flexible chain model, however, we used the guiding function

$$G(x) = E_{\mathrm{LJ}}(C_{\mathrm{HP}}; x). \qquad (15)$$

$G(x)$ omits the chain bending energy E'_{chain} and allows for a modified value of the Lennard-Jones parameter C_{HP}. The negative value of C_{HP} in the model has the effect that during gradient descent the condensed monomers may fission into separated H and P domains. This is avoided by giving C_{HP} a non-negative value in the guiding function.

III. RESULTS

A. Rotamer model

A useful record of the progress of the difference map algorithm is the time series of difference magnitudes $\delta_t = \|\Delta(x_t)\|$. In our folding application, δ_t is the rms displacement (in units of the chain's bond length) of monomers in two configurations: one satisfying the chain constraint, the other satisfying the packing (energy) constraint. The algorithm terminates when $\delta_t = 0$, that is, when a valid polymer geometry is found with energy below the chosen target value E_0 (a "feasible solution"). Figure 5 shows a difference plot with $\beta = 1.2$ for the sequence $(\mathrm{HPP})_8$ in the rotamer model with geometry $\cos\alpha = 0.5$ and target energy $E_0 = -24$. The behavior of δ_t in the folding problem is typical of behavior observed in other applications [6]. The three stages of the solution process are evident in (i) a fast initial decay (not shown in Fig. 5) during convergence to the quasiattractor, (ii) steady-state fluctuations as the quasiattractor is searched, and (iii) a final (fast) decay to zero when the solution (a fixed point) is discovered. As in phase retrieval, the distribution of run times (total iterations) is exponential [5] and consistent with the interpretation of a very fast relaxation of the probability distribution on the quasiattractor. For the parameters given, the average number of iterations per solution was $I_{\mathrm{av}} = 7500$.

The feasible solutions found by the difference map for given target energies E_0 were refined by steepest descent minimization of the heteropolymer energy; the chain geometry was maintained by adding the penalty function (12) with a large multiplier. For each run of the algorithm we therefore obtain one locally minimized fold with energy guaranteed to be below E_0. In the example above, about half of the outputs had the same refined energy of -25.048 and structure (or enantiomorph). Since this is also the lowest energy obtained, we have good reason to believe this is the ground state. The structure, shown in Fig. 6, resembles a cut trefoil knot.

The most direct measure of the work performed by the algorithm is the average number of iterations per solution I_{av}, divided by the rate p_0 with which the lowest-energy fold (putative ground state) is obtained. This is a number that we expect to grow exponentially with the length of the polymer, and roughly corresponds to the number of conformations that must be sampled before one can claim to have discovered the ground state. For the example above, $I_{\mathrm{av}}/p_0 \approx 15\,000$. We repeated the above experiment with longer sequences having the same repeating motif. The size $N = 36$ is about the limit of where the ground state can be established with modest computing resources (a single processor). As argued below, it may be possible to exceed this limit for well-designed se-

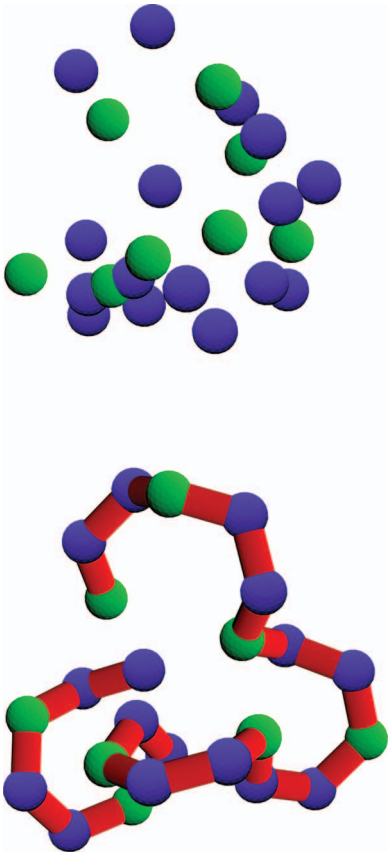

FIG. 4. (Color) Chain constraint projection applied to a typical monomer configuration (top) in the rotamer model. The projection (bottom) minimally displaces monomers in order to satisfy bond length and angle constraints.

quences. Our rotamer model experiments are summarized in Table I.

B. Flexible chain model

Studies of this model by other investigators [8,9] have been limited to Fibonacci sequences F_k, defined by

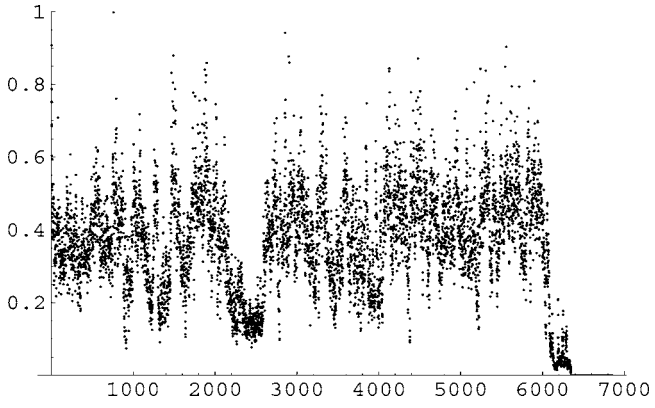

FIG. 5. Evolution of the rms displacement of monomers, δ_t, between two configurations that satisfy the chain and packing constraints, respectively. The fold shown in Fig. 6 below was found in just over 6000 iterations.

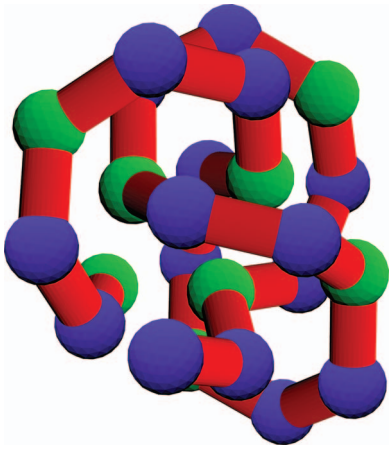

FIG. 6. (Color) The fold having the lowest energy for the sequence $(HPP)_8$ in the rotamer model has the shape of a cut trefoil knot. Hydrophobic monomers are colored green, polar monomers are blue.

$$F_0 = \mathrm{H}, \quad F_1 = \mathrm{P}, \quad F_{k+1} = F_{k-1}F_k. \tag{16}$$

The tendency toward hydrophobic core formation is even stronger for the Lennard-Jones parameters of the flexible chain model. For the Fibonacci sequences, in particular, the chain bending energy must be sacrificed in order to allow the chain to weave between the hydrophobic core and polar envelope. To improve the packing projection we therefore used the guiding function (15), which omits the bending energy, and $C_{\mathrm{HP}}=0$ for the short chains, $C_{\mathrm{HP}}=0.1$ for $N=55$. A sign change of the difference map parameter β, which effectively interchanges the two constraint sets, gave somewhat better results in the flexible chain model.

Our results for Fibonacci chains up to $N=55$ are summarized and compared with other algorithms in Table II. The difference map corroborates the ground state candidates found by the ELP [13] algorithm for chains up to $N=34$, and finds a lower-energy fold for $N=55$. All the best folds have a well developed hydrophobic core; the $N=55$ chain shown in Fig. 7 is a good example. The latter fold was only obtained in one run, and we are therefore far from claiming to have found the ground state.

Low-energy folds in the flexible chain model for sequences containing adjacent H monomers are qualitatively

TABLE I. Results for the rotamer model. E_0 is the target energy of the difference map (DM) algorithm, I_{av} the average number of iterations to find the target energy, and p_0 the probability that the discovered fold refines to the lowest energy obtained in the experiment, E_{DM}. The last column gives the CPU time per iteration on a 1.67 GHz processor.

N	Sequence	E_{DM}	E_0	β	I_{av}	p_0	Time/iteration (ms)
24	$(HPP)_8$	−25.048	−24.0	1.2	7500	0.50	12
30	$(HPP)_{10}$	−34.900	−33.0	1.2	23000	0.07	18
36	$(HPP)_{12}$	−45.851	−42.5	1.2	150000		26

TABLE II. Results for the flexible chain model. Ground state energy estimates obtained by the difference map (DM) algorithm are compared with three other algorithms: pruned-enriched Rosenbluth method (PERM [8]), multicanonical sampling (MUCA [9]), and energy landscape paving (ELP [9]). Optimal structures found by ELP and DM for $N=13$, 21, and 34 Fibonacci sequences are essentially the same [12]. For $N=55$ the DM finds a different, lower-energy fold. See Table I for definitions of DM parameters.

N	Sequence	E_{PERM}	E_{MUCA}	E_{ELP}	E_{DM}	E_0	β	I_{av}	p_0	Time/iteration (ms)
13	F_6	−4.962	−4.967	−4.967	−4.975	−4.5	−1	34	0.34	3
21	F_7	−11.524	−12.296	−12.316	−12.327	−11.8	−1	2900	0.024	25
34	F_8	−21.568	−25.321	−25.476	−25.512	−23.5	−1	10000	0.007	80
55	F_9	−32.884	−41.502	−42.428	−43.331	−38.0	−1	27000		200
25	$\mathrm{H(HPPH)_6}$				−28.313	−27.4	−1	9200	0.030	45

different from the low-energy folds for Fibonacci sequences, in which H monomers are never adjacent. A good example is provided by the $N=25$ sequence $\mathrm{H(HPPH)_6}$, which was designed to realize a particular ground state geometry. Using the difference map it is easy to establish the ground state shown in Fig. 8. This fold is unique in that it simultaneously minimizes the bending energy where the chain passes through the icosahedral core, and also arranges the six hairpin turns so that the P monomers there form the largest number of contacts.

IV. DISCUSSION

The difference map folding algorithm was shown to be competitive with leading algorithms in experiments with model proteins. We conclude by discussing the relationship of this algorithm to global optimization methods that have been applied to the folding problem, as well as two issues that will be important in applications to realistic protein models.

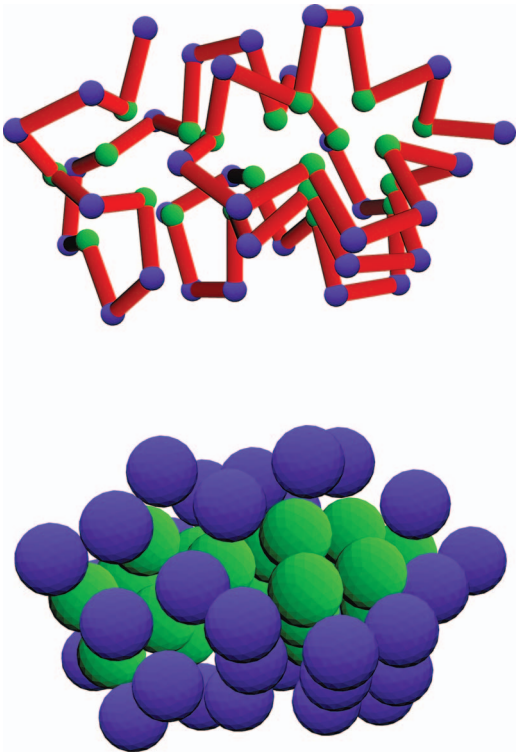

FIG. 7. (Color) Fold with lowest known energy for the $N=55$ Fibonacci sequence in the flexible chain model. Top, chain geometry; Bottom, monomer packing.

A. Relationship to global search

Algorithms that promise to find the true ground state of a protein model are useful not just for finding the native fold, but are essential in validating force field models of proteins in a solvent environment. Even at a rudimentary level of physical modeling, as in lattice-based HP models, global optimizers can shed light on the physical principles that lie at the core of protein structure and sequence design.

One of the most successful global search schemes is the constrained hydrophobic core construction (CHCC) method of Yue and Dill [14] and further developed by Backofen and Will [15]. This method has only been implemented on the lattice, where it can now find ground states and compute degeneracies for up to 200 HP residues. A weaker form of the core constraint has been applied to conformational searches of off-lattice models at the united atom level of detail [16,17].

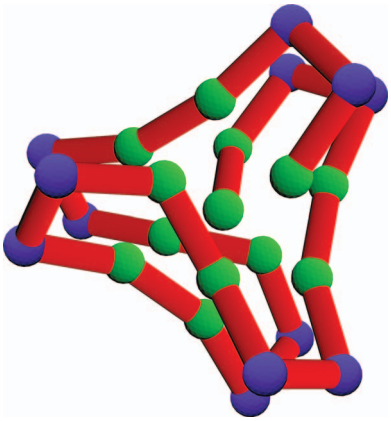

FIG. 8. (Color) Ground state of the designed sequence $\mathrm{H(HPPH)_6}$ in the flexible chain model. The 13 H monomers (green) form the vertices and center of an almost perfect icosahedron. Apart from the final bond in the structure, the chain geometry is approximately symmetric with respect to a twofold axis.

The CHCC is similar to our constraint-based method in that the tasks of forming a condensed core (packing constraint) and threading the sequence (chain constraint) are performed separately. However, unlike the difference map which performs these two tasks very much in tandem, the CHCC gives higher priority to the packing problem and tackles this first using a "greedy" strategy. CHCC first finds optimal and near optimal hydrophobic cores by minimizing surface area, the idea being that threading will probably succeed with a core that is not too far from optimal.

The corresponding task in the difference map algorithm is the projection to the packing constraint. The constraint set for this projection, fixed by the target energy E_0, is usually far from the optimal monomer packing energy, apparently because the chain constraint in the off-lattice setting plays a more significant role. Whatever the actual origin, it appears that the determination of the native fold is more equally shared between these competing constraints in the models of our study. Perhaps this can be interpreted even further, as a mechanism responsible for the diversity of protein folds. That is, one would expect a smaller diversity in shape if optimal or near-optimal packings were able to accommodate any sequence.

The difference map appears to work best when the two constraint sets that define the solution are competitive, and the computational workload is equally shared between the two constraint projections. The algorithm's chief drawback, relative to global searches such as CHCC, is that the property of being exhaustive rests on unproven ergodicity assumptions about the chaotic dynamics. However, this criticism applies also to phase retrieval, where iterative, constraint-based algorithms have no rivals.

B. Designed sequences

The performance of an iterative phase retrieval algorithm, of which the difference map folding algorithm is a logical descendent, is sensitively dependent on the degree to which the input data is overdetermined [6]. We believe that the latter attribute's counterpart in protein folding is the property of being well designed.

In the context of the geometry of the difference map, a highly overdetermined problem corresponds to the situation where the probability of nonempty intersection of the constraint sets A and B, given a specification by random data, is exceedingly small. This makes the existence of a solution all the more unusual. In phase retrieval one is guaranteed a solution in even these unlikely circumstances, and moreover, the uniqueness of the solution and efficiency of the solution process relies on this fact.

Whether the simple protein models studied above have the capacity for realizing highly overdetermined problem instances (sequences) is open to speculation. With our choice of deconstructing the energy landscape into chain and packing constraints, this would imply the existence of exceptionally low-energy monomer packings that nevertheless can be threaded by a particular sequence. Folds with these properties should be easier to find, because the target energy E_0 of the difference map algorithm could be set at a lower value and thereby eliminate a large part of the energy landscape. One experiment to test this hypothesis, in the flexible chain model, would be to fold random sequences of 13 H and 12 P monomers and compare performance, as well as ground state energies, with the designed sequence $H(HPPH)_6$.

C. More realistic models

A serious deficiency of the heteroploymer models studied above is the omission of the hydrogen bonding mechanism that acts on the peptide geometry and is responsible for the two distinctive types of secondary structure. Another deficiency is the neglect of side chain geometry and its effects on packing. The computational overhead resulting from these rather significant refinements would appear to present a daunting challenge, given that the task of finding ground states for the much simpler heteropolymer models seems insurmountable even for relatively short chains.

On the other hand, it may prove that the added level of detail in a realistic protein model is essential in the construction of well-designed folders, that the exceptional characteristics of evolved proteins cannot be realized without the benefit of secondary structure elements and optimized side chain packing. Since the difference map algorithm can directly exploit the designed (overdetermined) characteristics of an energy landscape, there is optimism that its success with realistic models may exceed, in terms of chain lengths, what can be achieved with the heteropolymer models considered here.

ACKNOWLEDGMENTS

V.E. thanks Ron Elber for discussions. This work was supported by NSF Grant No. DMR-0426568 and GAANN Award No. P200A030111 from the U.S. Department of Education.

[1] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[2] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).
[3] G. M. Sheldrick, Acta Crystallogr., Sect. A: Found. Crystallogr. **60**, s5 (2004).
[4] C. M. Weeks and R. Miller, J. Appl. Crystallogr. **32**, 120 (1999).
[5] V. Elser, Acta Crystallogr., Sect. A: Found. Crystallogr. **59**, 201 (2003).
[6] V. Elser, J. Opt. Soc. Am. A **20**, 40 (2003).
[7] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993).
[8] H. P. Hsu, V. Mehra, and P. Grassberger, Phys. Rev. E **68**, 037703 (2003).
[9] M. Bachmann, H. Arkin, and W. Janke, Phys. Rev. E **71**, 031906 (2005).
[10] V. Elser, J. Phys. A **36**, 2995 (2003).

[11] M. Levitt, J. Mol. Biol. **170**, 723 (1983).
[12] M. Bachmann (private communication).
[13] U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. **88**, 068105 (2002).
[14] K. Yue and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 146 (1995).
[15] R. Backofen and S. Will, J. Constraints (to be published).
[16] K. Yue and K. A. Dill, Protein Sci. **5**, 254 (1996).
[17] K. Ishikawa, K. Yue, and K. A. Dill, Protein Sci. **8**, 716 (1999).