Generalized Pattern Search Algorithm for Peptide Structure Prediction

Giuseppe Nicosia and Giovanni Stracquadanio
Department of Mathematics and Computer Science, University of Catania, Catania, Italy

ABSTRACT Finding the near-native structure of a protein is one of the most important open problems in structural biology and biological physics. The problem becomes dramatically more difficult when a given protein has no regular secondary structure or it does not show a fold similar to structures already known. This situation occurs frequently when we need to predict the tertiary structure of small molecules, called peptides. In this research work, we propose a new ab initio algorithm, the generalized pattern search algorithm, based on the well-known class of Search-and-Poll algorithms. We performed an extensive set of simulations over a well-known set of 44 peptides to investigate the robustness and reliability of the proposed algorithm, and we compared the peptide conformation with a state-of-the-art algorithm for peptide structure prediction known as PEPstr. In particular, we tested the algorithm on the instances proposed by the originators of PEPstr, to validate the proposed algorithm; the experimental results confirm that the generalized pattern search algorithm outperforms PEPstr by 21.17% in terms of average root mean-square deviation, RMSD C_α.

INTRODUCTION

When analyzing the complex structure of a biological system, proteins are the most attracting molecular devices. They are likely involved in all processes of a living organism; they are responsible for behavioral changes in the cells. Due to the important role of proteins in a biological system, molecular biologists are interested in looking for the function of each protein to understand how they can change the state and behavior of a cell and, possibly, to use their functions to treat diseases with specific drugs.

A fundamental feature determining the function of a protein is its three-dimensional structure, also known as tertiary structure. Therefore, understanding how the proteins organize themselves in three dimensions has a central role in discovering, understanding, and treating diseases. At this point, it is obvious that a reliable method is necessary that could help us to predict the three-dimensional structure of a protein.

There are many chemical approaches to determine the structure of a protein. Historically, the first one was the x-ray crystallography (1), which appeared in 1934, when Bernal and Crowfoot took the first x-ray photograph of a crystalline globular protein. Later, in 1980, Wuttrich introduced nuclear magnetic resonance (NMR) (2). Although both of these techniques are reliable, they have many drawbacks concerning the long period of time required to obtain a complete definition of a structure, and the high costs required. More specifically, x-ray crystallography can be applied only if it is possible to crystallize a protein into a regular lattice, whereas NMR works with proteins in a solute environment. This causes the protein to take different conformations, leading to many difficulties in determining a single good protein model.

There have been many efforts in determining the tertiary structure of a protein by using computational methods; they are very attractive because they can provide meaningful prediction at a fraction of the cost and time of the non-computational approaches. The computational way is very interesting, but there are two main problems to be taken into account: the first is how to formalize the protein structure prediction (PSP) problem in a manner suitable for the application of a given algorithmic method; and the second refers to the choice of a suitable algorithm to face the PSP problem.

The most common algorithms for protein structure prediction are centered on the thermodynamical hypothesis, which postulates that the native state of a protein is the state with the lowest energy value under physiological conditions. In general, this state corresponds to the lowest basins of a given energy surface (3).

Since the interactions comprising the energy function are highly nonconvex, the PSP can be tackled as a global optimization problem and, in particular as a minimization problem. According to Levinthal's paradox, an exhaustive search algorithm would take the present age of the Universe for a protein to explore all possible configurations and locate the one with the minimum energy (4).

Generally speaking, we can define a global optimization problem as

$$\min f(x),\ x \in \Omega \subseteq X, \tag{1}$$

where $X = \{x \in \mathbb{R}^n | x_L \leq x \leq x^U\}$ and $\Omega = \{x \in X | C(x)\}$, f is the objective function and C: $X \rightarrow R$ is the constraint function. In particular, the general definition is useful if we are interested in solutions that are not necessarily feasible, but are allowed to violate the constraints. In this case, we want to find the solution with the best objective function value that minimizes the constraint violation. However, if we want to treat only feasible solutions, the constraint function is defined

Submitted October 18, 2007, and accepted for publication March 20, 2008.

Address reprint requests to Giovanni Stracquadanio, Tel.: 39-095-738-3079; E-mail: stracquadanio@dmi.unict.it.

Editor: Kathleen B. Hall.

0006-3495/08/09/4988/12 $2.00

doi: 10.1529/biophysj.107.124016

as C: $X \rightarrow [T, F]$ and we can call it oracular constraint; it simply states whether a solution is feasible or not.

At this stage, there are three classes of computational approaches: homology modeling (5,6); threading (7,8); and ab initio (9,10–17). The first method tries to predict the structure starting from an experimental existing one with a significant sequence similarity to the target protein. The second, threading, tries to fit the target sequence to an experimentally similar fold when sequence homology between target and experimental structure is weak. The third, the ab initio approach, is the most interesting and challenging for computer scientists and molecular biologists: given a protein sequence and an energy function, we use an algorithm A, which produces the corresponding coordinates of all atoms for the given protein sequence, to find the protein conformation with the lowest possible potential energy.

Nowadays, these methods are the only useful ones when the fold to be predicted is totally unknown. This condition is frequently verified when we try to predict the tertiary structure of a peptide; peptides are small proteins of length ranging between 5 and 20 amino acids that control many functions of a living organism. In particular, there is a great number of bioactive peptides, and the determination of their three-dimensional structure is crucial for the production of specific drugs. Although the secondary structure, rather than the tertiary structure, is the principal factor affecting the binding properties of a peptide, they are less defined in these small molecules.

In our research, we introduce a new ab initio method based on the well-known class of generalized pattern search algorithms (Gps) (18–20) for the peptide structure prediction problem. Gps algorithms have a robust theoretical background and they have been successfully applied in several real-world applications (21–23). According to the thermodynamical hypothesis, we use Gps to minimize the empirical conformation energy program for peptides (ECEPP/3) potential energy function (24), a well-established potential energy function, to find the most plausible structure for a protein sequence.

Firstly, we outline some of the state-of-the-art algorithms in PSP; then we describe the Gps algorithm and we show a few results from both theoretical and practical points of view. Next, we outline the coding scheme, the adopted potential energy model, and the settings of the algorithm in our experimental protocol. Finally, we show the obtained results on a well-known set of proteins.

Computational methods in protein structure prediction

Due to the great impact of proteins in every field of biology, in the last 20 years many computational methods have been proposed to find the near-native tertiary structure of a protein. The protein structure prediction (PSP) is so challenging that, from 1992, an ad hoc competition called Critical Assessment of Techniques for Protein Structure Prediction (CASP) was created to evaluate the current state of art algorithms for PSP.

Actually, one of the best prediction methods is I-TASSER (25), also known as Zhang-Server, a multistep algorithm that combines homology, threading, and ab initio methods. In the first step, it simply threads the target sequence against a nonredundant Protein Data Bank (PDB) library to find global structure templates; subsequently, the templates are reassembled using the TASSER Monte Carlo algorithm, where predictions are recombined using additional information coming from the predicted accessible surface area and from the predicted secondary structure information. After clustering all predictions, the centroids are refined by choosing the conformation with the minimum energy.

An emerging and powerful method for the prediction of protein structures is the meta-server approach. This strategy tries to find good structures combining the output of a certain number of methods. In this class, LOMETS (26) is one of the best methods; it combines the output of nine of the most used algorithms in the literature (i.e., FUGUE (5); PROSPECT2 (27); SPARKS2 (28); SP3 (29); SAM-T02 (30); HHSEARCH (31); PPA1 (26); PPA2 (26); and PAINT (26)). The Robetta server (14) combines homology modeling and de novo tertiary structure prediction with Ginzu homology identification and with a domain parsing protocol to provide prediction for the full length of each target. In the homology step, the algorithm combines consensus score with energetic selection from a model ensemble; the model ensembles are parametrically generated using K*Sync (32) for the alignment method for the template regions and the Rosetta (33) modeling loop for unaligned regions. Moreover, the loop regions are based on the generation of a large number of decoys using the Rosetta fragment assembly protocol. The filtered ensemble is structurally clustered, and the top five clusters are returned to build the final prediction. Side chains are added using a backbone-dependent rotamer library (34) with a Monte Carlo conformational search procedure.

An alternative algorithm is Raptor (35), based on the mathematical theory of linear programming (LP). It tackles the PSP using the threading approach, and it formalizes the protein-threading problem as an LP problem. The main advantage is the opportunity of using existing powerful LP algorithms to predict the tertiary structure. At the late CASP competition, when Raptor was presented in an enhanced version, its ability of producing high quality solutions was confirmed.

In the field of peptide structure prediction, to our knowledge, the most effective methods are PEPstr (35), PepLook (37), and Robetta (14). PEPstr starts from the observation that β-turn is an important and consistent feature of small peptides in addition to the regular secondary structures. In particular, it combines regular secondary structures and β-turn, and generates four models for each peptide: the first one models the peptide in extended conformation ($\phi = \psi = 180°$); the second one uses constrained conformations

derived from secondary structure information; the third one extends the second model by introducing β-turn information; and the last one extends the third model by assigning χ-angles on the basis of the Dunbrack rotamer library (34). All these models are subject to energy minimization using the Assisted Model Building and Energy Refinement (AMBER) Ver. 6 (38).

Generalized pattern search algorithm for nonlinear optimization

Generalized pattern search algorithms were defined and analyzed by Lewis and Torczon (20) for derivative-free unconstrained optimization on continuously differentiable functions, and they later extended them to bound constrained optimization problems.

The Gps for unconstrained or linearly constrained minimization generates a sequence of iterates $\{x_k\}$ in $\mathbb{R}^n$ with nonincreasing objective function values. Each iteration is divided in two phases: the Search phase and the Poll phase. In the Search phase, the objective function is evaluated at a finite number of points on a mesh. Formally, we define a mesh as a discrete subset of $\mathbb{R}^n$ where the fineness is parameterized by the mesh size parameter $\Delta_h > 0$. The main task of the Search phase is to find a new point that has a lower objective function value than the best current solution, called the incumbent. At least from a theoretical point of view, any strategy may be used to select the mesh points that are candidates to replace the incumbent. When the incumbent is replaced, i.e., $f(x_{k+1}) < f(x_k)$, then x_{k+1} is said to be an improved mesh point. Starting from this consideration, we can introduce a Search procedure based on surrogates (40,41); we can formalize a surrogate model of the given problem by tackling the optimization of the surrogate function using some derivative-based optimization tool or some quadratic programming procedure, and then moving the solution to a nearby mesh point, hoping to obtain a better next iterate (40). This is the approach used in the Boeing Design Explorer software (21).

When the Search step fails to provide an improved mesh point, the algorithm calls the Poll procedure. This phase consists in evaluating the objective function at the neighboring mesh points, to see whether a lower objective function value can be found. When the Poll fails to provide an improved mesh point, the current incumbent solution is then said to be a local mesh optimizer. When the algorithm finds a local mesh optimizer, it refines the mesh by using the mesh size parameter of

$$\Delta_{k+1} = \tau^{w_k}\Delta_k, \tag{2}$$

where $0 < \tau^{w_k} < 1$, and $\tau > 1$ is a real number that remains constant over all iterations, and $w_k \leq -1$ is an integer bounded below by the constant $w \leq -1$. When either the Search or the Poll steps produce an improved mesh point, the current iteration stops and the mesh size parameter may be kept constant or increased according to Eq. 2, but with $\tau > 1$ and with $w_k \geq 0$ being an integer that is bounded above by $w^+ \geq 0$. Using the previous equation, it follows that for any $k \geq 0$, an integer $r_k \in \mathbb{Z}$ exists such that

$$\Delta_{k+1} = \tau^{r_k}\Delta_0. \tag{3}$$

The basic element in the formal definition of a mesh is the set of positive spanning directions $D \in \mathbb{R}^n$; in particular, nonnegative linear combinations of the elements of the set D span $\mathbb{R}^n$. The directions can be chosen using any strategy, but this must assure that each direction $d_j \in D$, $\forall j = 1, 2, \ldots, |D|$, is the product $G\bar{z}_j$ of the nonsingular generating matrix $G \in \mathbb{R}^{n\times n}$ by an integer vector $\bar{z} \in \mathbb{Z}^n$; it is important to recall that the same matrix is used for all directions. We let D denote a real valued matrix $n \times |D|$, and similarly, $\bar{Z}$ denotes the matrix whose columns are $\bar{z}_j, \forall j = 1, \ldots, |D|$; at this point we can define $D = G\bar{Z}$. Using the Poll procedure, the mesh is centered around the current iterate $x_k \in \mathbb{R}^n$ and its fineness is parameterized through the mesh size parameter Δ_k as

$$M_k = \left\{x_k + \Delta_k Dz : z \in \mathbb{Z}_+^{|D|}\right\}, \tag{4}$$

where $\mathbb{Z}_+$ is the set of nonnegative integers. At each iteration, some positive spanning matrix D_k composed of the columns of D is used to construct the Poll Set. This consists of the mesh points neighboring the current iterate x_k in the directions of the columns of D_k, as in the following equation:

$$\text{Mesh points} = \{x_k + \Delta_k d : d \in D_k\}. \tag{5}$$

Theoretical results and real-world applications

In the case of bounded constraint optimization, Audet and Dennis (19) prove that if there is a convergent subsequence of the sequence $\{x_k\}$ of iterates produced by the algorithm (since $\{f(x_k)\}$ is nonincreasing), then it is convergent to a finite limit if it is bounded below. So, if f is lower semicontinuously at any limit point $\bar{x}$ of the sequence of iterates, then $f(\bar{x}) \leq \liminf_k f(x_k) = \lim_k f(x_k)$. Moreover, they show that there is a limit point $\hat{x}$ of a subsequence of $\{x_k\}$ consisting of iterates on progressively finer meshes; these specific iterates of interest are mesh local optimizers in that they minimize the function on a positive spanning set of neighboring mesh points. The directional tests that led Gps to refine the mesh at mesh local optimizers are exactly the difference quotients that are nonnegative for the Clarke generalized directional derivative $\hat{x}$. If the Clarke derivatives exist at $\hat{x}$, as they will if f is locally Lipschitz at $\hat{x}$, then these nonnegative difference quotients pass through the limit to be nonnegative Clarke derivatives in the used direction. It is clear that nonnegative directional derivatives in a set of directions are necessary conditions for optimality, but they are not the usual first-order conditions; to match them, it is assumed that the generalized gradient of f is a singleton. This

constraint causes the directional optimality conditions to hold for all directions in their positive cone and, with a right strategy for choosing directions, it leads to the first-order optimality conditions.

In addition to the theoretical results, Gps has been largely applied to a large number of real problems. Zhao et al. (42) have recently applied pattern search methods for the determination of a surface structure of nanomaterials (42); in particular, Gps has been used to fit low energy electron diffraction data with the experimental data. Although the problem is very hard, due to the presence of many local minima, Gps works better than the other state of art algorithms. Allison et al. (22) have applied Gps algorithms to aircraft design; they developed a decomposition-based method that is applied on the modeling of the various parts of an aircraft that share similar components. To this purpose, Gps was largely applied as the main optimization algorithm. Abramson (23) applied Gps for the optimization of a load-bearing thermal insulation system (which is characterized by hot and cold surfaces with a series of heat intercepts and insulators between them). The optimization problem is represented as a mixed variable programming problem with nonlinear constraints, in which the objective is to minimize the power required to maintain the heat intercepts at fixed temperatures so that one surface is kept sufficiently cold. In many of the faced real-world applications, Gps outperforms the corresponding state-of-the-art optimization algorithms.

METHODS

In this section, we report our main choices about the representation of protein conformations, the adopted energy function, and the metrics used to assess the structural qualities of the best protein conformations.

Coding conformations

A nontrivial task that precedes use of any optimization algorithm to tackle the PSP is the selection of a good representation for the protein conformations. The packing of amino acids produces a so-called polypeptide chain, where the backbone atoms are linked through the peptide bond. The fold of peptides can be described by using angles of internal rotations in the main chain. Internal rotations around N and C_α atoms, and C_α and C atoms are not restricted by the electronic structure of the bond, but only by possible steric collisions in the conformations. The side-chain conformations can be expressed by using angles of internal rotation, denoted by $\chi_1,\ldots, \chi_n$; the conformation of any side chain corresponding to different combinations of values of χ-angles are called rotamers. In the current work, we use an internal coordinates representation (torsion angles), which is currently the most widely used representation model. Each residue type requires a fixed number of torsion angles to fix the three-dimensional coordinates of all atoms. Bond lengths and angles are fixed at their ideal values.

In all simulations, all the ω-torsion angles are fixed, so the degrees of freedom of the representation are the main-chain and side-chain torsion angles (ϕ, ψ, and χ_i). The number of χ angles depends on the residue type, and they are constrained in regions derived from the backbone-independent rotamer library (34). Side-chain constraint regions are of the form: $[\mu - \sigma, \mu + \sigma]$, where μ and σ are the mean and the standard deviation, respectively, for each side-chain torsion angle computed from the rotamer library. It is important to note that, under these constraints, the conformation is still highly flexible and the structure can take on infinite various shapes that are vastly different from the native shape. In protein structure prediction it is crucial to know the existence of regular secondary structures; from this information, it is possible to set tighter bounds on the dihedral angles, which is useful to guide the algorithm to feasible and high quality solutions.

The secondary structure information, when used, was predicted by the Scratch prediction server (44), and the relative bounds for the main-chain dihedral angles are set according to Klepeis and Floudas (16), whereas the ω-angle is fixed to 180°.

Potential energy model

The interactions of the side chains and main chains with each other, with the solvent and with the ligands, determine the energy of the given protein conformation. The folding of a protein is a process that drives the atoms to be stabilized into a conformation that is better than others, the so-called native state. The formation of the native state is a global property of a protein, because the stabilizing interactions involve parts of the protein that are distant in the polypeptide chain but near in space. In particular, Anfisen et al. (3) states the native state is the one with the lowest free energy. From a thermodynamic point of view, the free energy of a protein depends on the entropy and on the enthalpy of the system. Without losing generality, we can assume that a protein can only be in two states: folded and unfolded; at low temperature, the energy of the folded state is lower than that of the unfolded state. Since we are interested in the folded state of a protein, we consider as a good candidate structure the one with lowest energy. Under ordinary conditions, the free energy of the stabilization of proteins is typically in the range 5–15 kcal/mol (45).

It is clear that computing the free energy of a system in silico is impossible, because we are not able to simulate complex chemical systems that mutate in the time. So we need an analytical expression that gives information about the thermodynamical state of a protein as a function of the position of the atoms; this is the so-called potential energy function. Most typical potential energy functions have the form

$$E(\vec{R}) = \sum_{\text{bonds}} B(R) + \sum_{\text{angles}} A(R) + \sum_{\text{torsions}} T(R) + \sum_{\text{nonbonded}} N(R), \quad (6)$$

where $\vec{R}$ is the vector representing the conformation of the molecule, typically in Cartesian coordinates or in torsion angles.

The first three terms describe the local interactions between atoms that are separated by one, two or three covalent bonds; many proteins contain covalent bonds in addition to those of the polypeptide backbone and of the side chain. In particular, the first term refers to the bond length stretching, the second one to the angle bending, and the last one represents the angle twisting. The last term takes into account the nonlocal interactions between pairs of atoms that are separated along the covalent structure by at least three bonds. In particular, one of the main nonbonded actors are the van der Waals forces; the packing of atoms in a protein contributes to the stability of the protein itself by excluding the nonpolar atoms from contact with water and by packing together the atoms of the protein. The literature on proper cost functions is enormous (46–49). In this work, we use the empirical conformation energy program for peptides (ECEPP) potential energy function version 3 (24). In this model, the lengths of covalent bonds, along with the bond angles, are taken to be constant at their equilibrium value, and the independent degrees of freedom become the torsional angles of the system. The potential energy function E_{tot} is the sum of the electrostatic term E_c, Lennard-Jones term E_{LJ}, and the hydrogen-bonding term E_{HB} for all pairs of peptides, together with the torsion term E_{tor} for all torsion angles. The function has the form

$$E_{\text{tot}} = E_C + E_{LJ} + E_{HB} + E_{\text{tor}}, \quad (7)$$

$$E_C = \sum_{(i,j)} \frac{q_i q_j}{r_{ij}}, \quad (8)$$

$$E_{\mathrm{LJ}} = \sum_{(\mathrm{i,j})} \left(\frac{A_{\mathrm{ij}}}{r_{\mathrm{ij}}^{12}} - \frac{B_{\mathrm{ij}}}{r_{\mathrm{ij}}^{6}} \right), \tag{9}$$

$$E_{\mathrm{HB}} = \sum_{(\mathrm{i,j})} \left(\frac{C_{\mathrm{ij}}}{r_{\mathrm{ij}}^{12}} - \frac{D_{\mathrm{ij}}}{r_{\mathrm{ij}}^{10}} \right), \tag{10}$$

$$E_{\mathrm{tor}} = \sum_{\mathrm{l}} U_{\mathrm{l}}(1 \pm \cos(\eta_{\mathrm{l}}\chi_{\mathrm{l}})). \tag{11}$$

In this model, r_{ij} is the distance between atoms i and j, and χ_{l} is the torsion angle for chemical bond l. The bond lengths and bond angles (which are hard degrees of freedom) are fixed at experimental values, and dihedral angles ϕ, ψ, ω, and χ_{i} are independent variables. The various parameters (q_{i}, A_{ij}, B_{ij}, C_{ij}, D_{ij}, U_{l}, and η_{l}) were determined by a combination of a priori calculations and minimization of the potential energies of the crystal lattices of single amino acids. As already stated, the free energy of folding of a protein consists of the sum of contributions from the energy of its intramolecular interactions and from the free energy of interaction of the molecule with the surrounding solvent water; however, exact computation of the solvent contribution is very complex (50–56).

In this study, we use the model proposed by Ooi et al. (57): they assume that the extent of interaction of any functional group i of a solute with the solvent is proportional to the solvent-accessible surface area A_{i} of group i because the group may interact directly only with the group at this surface. The total free energy of hydration of a solute molecule is given by

$$\Delta G_{\mathrm{h}}^{\mathrm{o}} = \sum_{\mathrm{i}} g_{\mathrm{i}} A_{\mathrm{i}}, \tag{12}$$

where the summation extends over all groups of the solute and A_{i} is the conformation-dependent accessible surface area of group i, whereas the constant of proportionality g_{i} represents the contribution to the free energy of hydration of group i per unit-accessible area.

The model is adopted because it is specifically designed to supplement the ECEPP algorithm. The free energy of hydration, to be added to the ECEPP energy, must correspond only to the additional interactions of the atoms of the solute with water.

All the potential energy calculations have been conducted using the Simple Molecular Mechanics for Proteins (SMMP) (58), which is a Fortran package designed for molecular simulation of linear peptides.

Algorithm settings

Gps was tested using the settings reported in Table 1. The Latin hypercube sampling (59) has been chosen as a method to sample the space of solutions.

TABLE 1 Gps parameters

Settings	Gps
Initial poll size	1
Max poll size	128
Poll directions	$2 \times n$
Coarsening exponent	1
Refining exponent	-1
Search strategy	Latin hypercube
Initial search generated points	$n \times 100$
Iterative search generated points	$n \times 2$

Fixed n as the number of dihedral angles for a given protein, the initial poll size is the initial step length of the algorithm, whereas the max poll size is the maximum allowed step length; poll directions are defined using the coordinate search; coarsening and refining exponents are used to increase the step length due to successful or unsuccessful iteration; the initial and iterative search generated points are the points generated during the initial sampling and the poll phase, respectively.

In our work, two types of Search phase have been used: the initial search and the iterative search. The first one uses the search procedure to explore the landscape of solutions starting from a given initial point; the best point, in terms of potential energy value, is kept as the initial point for the Gps main loop as described previously. The second type, the iterative search, is performed during the Gps loop. The number of points generated by the initial search was set to $n \times 100$, where n is the number of dihedral angles, because it is crucial to find a good starting point for Gps. In the iterative phase, however, it was set to $n \times 2$, because it has been shown that the pattern search phase becomes more effective than the search procedure (60). Moreover, we use a coarsening exponent fixed to 1, and the refining exponent fixed to -1, to prevent a rapid convergence of the algorithms and to avoid the possibility of getting trapped in local optima. The maximum poll size (p_{s}) is the longest step length that the algorithm can perform; we set it to 2^7, because it is a good tradeoff between the probability of making huge jumps in the solution space and the probability of minimizing the number of discarded solutions due to the filter constraints approach. The stopping criterion was fixed to the achievement of a mesh with fineness 0.5; this is justified because we want to consider only integer dihedral angles.

Metrics

In our research work, to measure the quality of the solutions found, we use a well-known measure: the root mean-square deviation (RMSD) measured on all atoms and on C_α atoms of our best solution found against the corresponding structure stored in PDB. Moreover, for each conformation we report the free energy value measured in kcal/mol. Since the RMSD weighs the distances between all residue pairs equally, a small number of local structural deviations could result in a high RMSD, even when the global topologies of the compared structures are similar. Moreover, the average RMSD of randomly related proteins depends on the length of compared structures, which renders the absolute magnitude of RMSD meaningless. For this reason, the CASP competition has focused its attention on the necessity of a reliable and effective metric to assess the quality of the predicted structures; actually, in CASP7, three metrics are used to evaluate the quality of the solution found: GDT (61), MaxSub (62), and TM-score (63).

GDT tries to identify any accurately, and not necessarily contiguous, predicted substructures. This metric attempts to find the maximum number of predicted residues that can be superimposed over the reference structure within a given threshold. Unfortunately, the task of finding the largest subset of residues superposed at a given threshold is a hard problem, hence approximations need to be used.

MaxSub exploits the principles of GDT, giving a more accurate measure. The returned value is a normalization of the size of the largest well-predicted subset and is computed using a variation of a formula suggested by Yona and Levitt (64). Formally, given two ordered sets of points in a three-dimensional space, $A = \{a_1, a_2, \ldots, a_{\mathrm{n}}\}$ and $B = \{b_1, b_2, \ldots, b_{\mathrm{n}}\}$, where A is the reference structure and B is the experimental one: for each residue i, a_{i}, and b_{i} are the corresponding three-dimensional coordinates. We can define a match as an ordered set such that $M = \{(a_{\mathrm{i}}, b_{\mathrm{i}}) | a_{\mathrm{i}} \in A, b_{\mathrm{i}} \in B\}$, where $|M| \leq n$. A match defines an optimal transformation T that best superimposes the points of B over A, such that T minimizes:

$$RMS(M) = \sqrt{\frac{\sum_{(\mathrm{a_i, b_i}) \in \mathrm{M}} \| a_{\mathrm{i}} - T(b_{\mathrm{i}}) \|^2}{|M|}}. \tag{13}$$

Here $\|\bullet\|^2$ is the Cartesian distance. The MaxSub score tries to find the largest subset M such that $\|a_{\mathrm{i}} - T(b_{\mathrm{i}})\|^2$ is below some threshold; it is largely accepted to set this threshold to 3.5 Å.

The recently proposed TM-score (Eq.14) overcomes these problems by exploiting a variation of the Yona and Levitt (64) weight factor that weighs the residue pairs so that those at smaller distances are relatively stronger than those at larger distances. Therefore, the TM-score is more sensitive to the global topology than to the local structural variations. It follows the TM-score definition

$$TM\text{-}score = Max\left[\frac{1}{L_N}\sum_{i=1}^{L_r}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right], \qquad (14)$$

where L_N is the length of the native structure, L_r is the length of the residues aligned to the reference structure, d_i is the distance between the i^{th} pair of aligned residues, and d_0 is a scale to normalize the match difference. As denoted by the Max function, the TM-score is the maximum value after optimal spatial superposition.

RESULTS

In this section, we show the performances of Gps over a well-known set of peptides. Our experimental protocol is divided into two stages.

The first experimentation is concerned with the comparison of Gps with another state-of-the-art pattern search-based algorithm, known as mesh adaptive direct search (MADS) (60) and parallel pattern-search swarm (PPswarm) (65), on two classic testbeds for peptide structure prediction problem: the met-enkephalin (PDB Id.: 1PLW) and the melittin (PDB Id.: 2MLT).

In the second experimentation, we have tested and compared Gps with PEPstr on a set of 42 bioactive peptides proposed by Kaur et al. (36), which is the state-of-the-art for peptide structure prediction.

Met-enkephalin

The first peptide used in our experiments is the met-enkephalin (PDB Id.: 1PLW) (66). The met-enkephalin is a small peptide composed of five residues that occur naturally in human brain and in pituitary gland. This is the peptide (H-Tyr-Gly-Gly-Phe-Met-OH), which contains 75 atoms that define 24 dihedral angles. Despite the small length of this molecule, it has been estimated that $\sim 10^{11}$ distinct local minima of the potential energy function exists for this protein (10). Due to these features, this peptide has received a great attention for many optimization algorithms that try to find the native structure of a protein through the minimization of a potential energy function (9,10,67). The peptide does not define any regular secondary structure, so the bounds for main-chain dihedral angles ϕ, ψ-angles were set to $-180° \leq \phi, \psi \leq 180°$, and $\omega = 180°$; the side-chain dihedral angles are constrained using a well-known rotamer library (34).

We performed 10 independent runs of the Gps algorithm starting from 10 different random conformations. The best conformation found has a potential energy of -42.918 kcal/mol, which reports an RMSD on C_α atoms of 0.961 Å: the superposition with the corresponding structure stored in the PDB is shown in Fig. 1. This prediction reports a TM-score of 0.5765, MaxSub of 0.9341, and GDT of 0.9500, which confirm the quality of the predicted structure. The measurements of all predicted structures using these metrics are presented in Fig. 2. Over the 10 runs performed, the mean energy conformation is located at -39.294 ± 2.37 kcal/mol; the algorithm reaches this local optimum with an average number of 173.8 iterations, using an average number of 10281 function evaluations. For this peptide the putative energy global minimum is -11.707 kcal/mol (24). Gps successfully locates this minimum after 384 objective function evaluations.

MADS and PPSwarm performs worse than Gps, as reported in Table 2. MADS reported a conformation with an energy value of -40.812 kcal/mol while PPSwarm found a conformation with potential energy value of -37.412 kcal/mol; the poor quality of the structure is confirmed in both cases by a high value of RMSD.

We studied the average amount of time needed by Gps to reach the stopping criterion. All the simulations were conducted on a Pentium IV 3.0 Ghz with 256 Mb of SDRAM, running a GNU/Linux Debian 3.1 operating system. The average time, measured using the Unix time command, is 12 min 32 s; we believe that we can improve this running time by developing a parallel Gps, where in the Poll phase we can adopt a classical parallelization scheme for the pattern search (68).

Clustering analysis

Starting from these results, we want to study how many distinct and locally optimal conformations have been found by the algorithm. To obtain this information, we conducted a cluster analysis of all the conformations produced in the best Gps run. The clustering takes only proteins with negative potential energy values into account, because a conformation with positive energy is considered infeasible. It is clear that

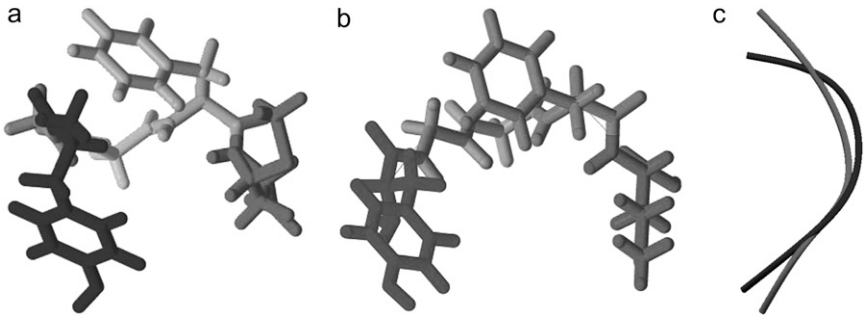

FIGURE 1 A comparison of the structure stored in the PDB for the met-enkephalin (*a*) peptide and the structure predicted by Gps (RMSD $C_\alpha = 0.981$ Å) (*b*) and the corresponding structure superposition (*c*), where, in light shading, we plot the GPS predicted structure and in dark shading, the corresponding PDB structure (PDB Id.: 1PLW).

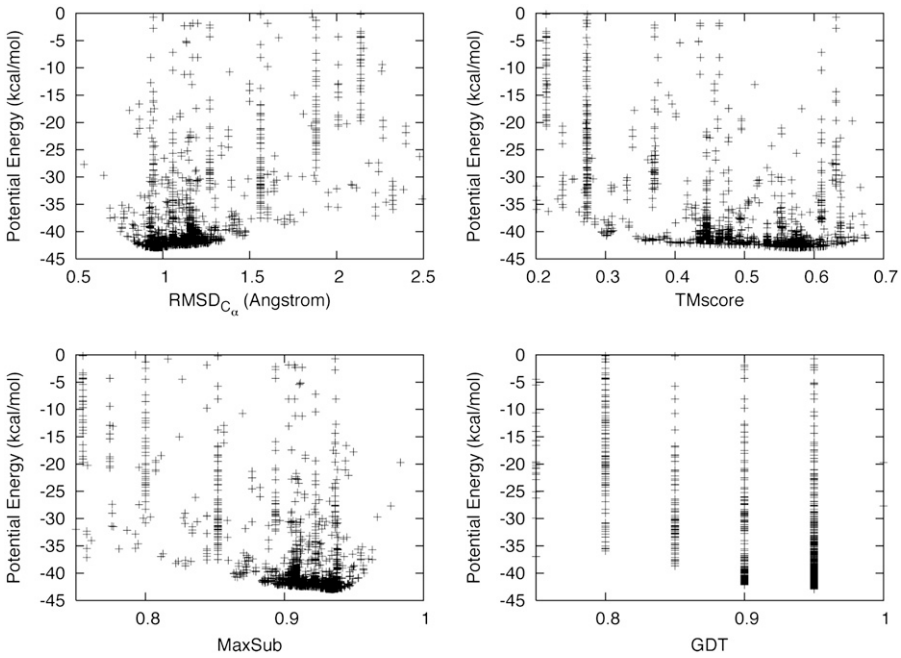

FIGURE 2 Evaluation of all conformations predicted by Gps for the met-enkephalin (PDB Id.: 1PLW) using RMSD C_α, TM-score, MaxSub, and GDT metrics.

we need a similarity function that could help us to group similar structures together. In our work we label each conformation according to the Zimmerman conformational code (69); this is a coding scheme that assigns a letter to each residue on the basis of the value of main-chain dihedral angles. In our work, for each conformation produced, we assign the relative Zimmerman code and we group together the proteins that have an equal code for the three central residues (9): as a representative member of each cluster, we choose the protein with the lowest value of the potential energy function. Out of 12,651 protein conformations predicted, 4486 are feasible proteins, grouped into 22 distinct clusters, and 8165 are infeasible proteins. In Table 3, we report the first five ranked clusters. For each of them, we report the Zimmerman code, the number of conformations that belong to it, the value of the potential energy of the best conformation in the cluster, and the corresponding RMSD C_α.

By analyzing the obtained clusters, one may observe that the biggest cluster is the one which contains the best solution, in terms of potential energy value, obtained by Gps. This is not surprising, primarily because Gps starts each exploratory move from a single point, the best found so far, and it tries to explore its neighborhood as deeply as possible hoping to find something better. Moreover, the analysis of the clusters reveals that the best conformation, in terms of RMSD C_α, is located in the fifth cluster, where the representative conformation has a potential energy of −37.756 kcal/mol; this result confirms that there is not a bijective correspondence between low potential energy value and good RMSD value.

TABLE 2 Comparison between pattern-search based algorithms on the met-enkephalin peptide (PDB Id. 1PLW)

Algorithm	Potential energy (kcal/mol)	RMSD C_α Å
Gps	−42.918	0.961
MADS	−40.812	3.278
PPSwarm	−37.412	3.422

For each algorithm we report the best solution in terms of potential energy, and its relative RMSD C_α.

Melittin

Melittin (PDB Id:2MLT) is a peptide of 26 amino acids that has recently received a good deal of attention in computational protein folding (9,70). In particular, the membrane portion of this protein has huge number of local minima, believed to range between 10^{34} and 10^{54} (70). The membrane-

TABLE 3 1PLW conformational clustering

Cluster rank	Zimmerman code	No. conformations	Potential energy (kcal/mol)	RMSD C_α Å
1	E*CA	4330	−42.918	0.961
2	E*HF	51	−20.673	2.010
3	E*DA	27	−38.171	1.237
4	H*CA	15	−39.741	1.364
5	E*BA	13	−37.576	0.776

For each conformation explored, we assign the relative Zimmerman code (69) and we group together the proteins that have equal code for the three central residues (Gly-Gly-Phe) (9); as a representative member of each cluster, we choose the protein with the lowest potential energy value. For each cluster, we report the number of conformations that belongs to the cluster, the potential energy function of the representative member, and its relative RMSD C_α.

bound portion of the protein is composed of 20 amino acids, and it defines 84 dihedral angles and 402 atoms.

This peptide has two α-helices connected by a small loop region; in this case, to study the impact of the constraints derived from secondary structure, we perform five runs using a fully extended representation without secondary structure information, and 10 runs using ad hoc constraints derived from the secondary structure as defined in Klepeis and Floudas (16). The best conformation found by Gps using the extended representation has potential energy of -80.817 kcal/mol, with a RMSD on C_α atoms of 5.8 Å and an average energy solution of -71.227 ± 8.678 kcal/mol. One may note that the predicted conformation is far from the native one, which is confirmed by the Zimmerman code for the 18 central residues (AC*CAEDACFCD*DDECEA*E), where it is observed that the two α-helices are not defined at all. By inspecting the results obtained using the constrained representation, we are able to predict a conformation with a potential energy of -104.349 kcal/mol, with an RMSD C_α of 3.089 Å (Fig. 3). It also reports a TM-score of 0.392, MaxSub of 0.553, and GDT of 0.725 (Fig. 4). This prediction requires 875 iterations and 83,514 function evaluations; and the average potential energy value over the 10 runs is -94.8134 ± 13.863 kcal/mol. By inspecting the solvation term of the energy function, which takes a value of -20.26 kcal/mol, one may note that the protein is well exposed to the solvent. The putative energy global minimum of this protein is -91.02 kcal/mol (9). By inspecting the ensemble of conformations predicted by Gps, this minimum has been located after 27,778 objective function evaluations, and its RMSD on C_α atoms is 3.329 Å.

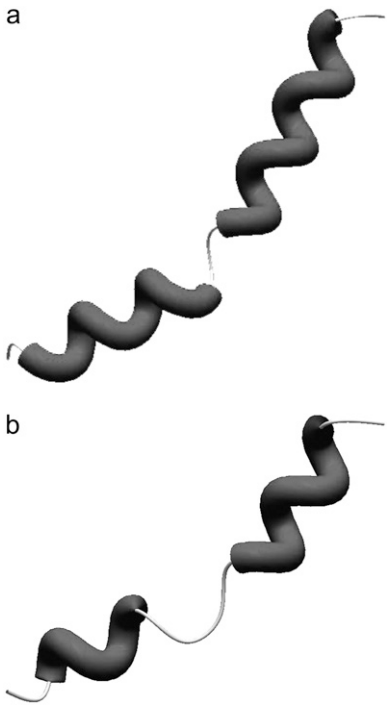

FIGURE 3 A comparison of the structure stored in the PDB for the melittin (*a*) peptide (PDB Id.: 2MLT) and the structure predicted by Gps (RMSD $C_\alpha = 3.0891$ Å) (*b*).

In Table 4, it is possible to note that, in this case, MADS and PPSwarm also perform worse than Gps, both with and without secondary structure information; MADS reported a conformation with potential energy of -78.124 kcal/mol without secondary structure information, and -94.780 kcal/mol with secondary structure. In all the experiments, PPSwarm does not find any feasible conformation, because all candidate solutions have a positive energy. In particular, it traps at 100.312 kcal/mol. The application of cluster analysis to all conformations predicted during the best run of Gps shows that, out of 83,514 conformations, only 53,212 are feasible conformations, and these are clustered into one group, labeled by the Zimmerman code AAAAAAAACDB*AAAAAAA. In Table 5, we report the dihedral angles of the conformation with the lowest potential energy value compared with the dihedral angles predicted by Klepeis et al. (9).

PEPstr benchmark

To prove the effectiveness of the Gps algorithm for the prediction of the three-dimensional structure of a peptide, we have conducted an extensive series of simulations on the same test bed proposed by Kaur et al. (36).

This benchmark is composed of 77 experimentally determined three-dimensional structures of bioactive peptides; only a few structures are solved using x-ray crystallography, and most of them have NMR-solved structures. From these 77 structures, the authors excluded 35 peptides stabilized by disulfide bridges.

The remaining set of 42 molecules can be grouped according to their regular secondary structure: 32.3% are α-helices, 6.9% are β-sheet, and the remaining 34.9% are β-turns.

The authors validate their algorithm, known as PEPstr, on this benchmark. They provide four models for each protein, where the first model is obtained by using an extended conformation; the second model by using constrained conformation for ϕ, ψ-bound based on the regular secondary structure information; the third model extends the second one by introducing β-turns information; and the last one adds side-chain angles from the rotamer library to the third model.

All these models undergo energy minimization and molecular dynamics calculations using SANDER module with the AMBER force field, the distance-dependent dielectric constant, and the nonbonded cutoff value of 8 Å.

In our experiments, we predict a single model, the one with the lowest potential energy, using no secondary structure information for peptides with <15 amino acids; we just use the extended conformation also when there are defined secondary structures. Moreover, for each instance, we perform five independent runs starting from random protein conformations. The results are reported in Tables 6–8, where we

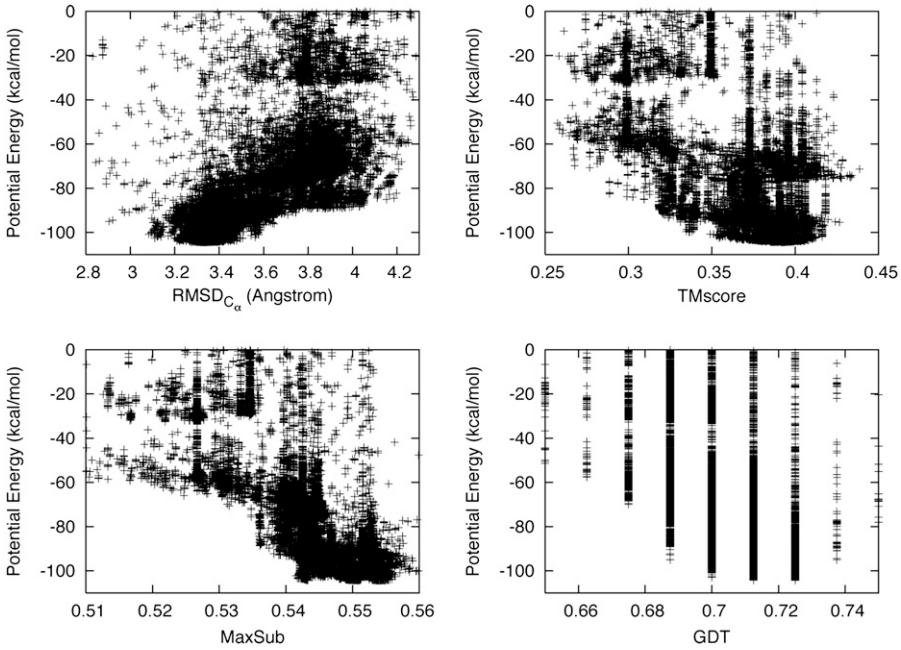

FIGURE 4 Evaluation of all conformations predicted by Gps for the melittin (PDB Id.: 2MLT) using RMSD C_α, TM-score, MaxSub, and GDT metrics.

show the best potential energy found, the relative van der Waals forces and solvent term, and finally the RMSD on C_α atoms against the corresponding structure stored in the PDB. One may note that all the structures have a negative potential energy value, and the negative contribution by the van der Waals forces assesses that they are feasible conformations. Moreover, all of them are well exposed to the solvent, as it is possible to infer from the value of the solvation term.

To compare Gps with PEPstr, we calculate the mean RMSD C_α on all the 42 instances whereby Gps reports a value of 3.153 Å as shown in Table 9; if we compare this result with the best performance obtained by PEPstr, using model IV, they report an average RMSD C_α of 4.0 Å, that is, worse than Gps by 21.17%. It is interesting to note that Gps performs quite well in all the instances, even for peptides that define coil regions or β-sheet (Fig. 5). Probably, with the addition of information on β-turns, and in general on the regular secondary structure information for all peptides, we can improve this performance, because the Poll phase works better with tighter constraints.

CONCLUSIONS

The prediction of the three-dimensional structure of a protein is one of the open problems in structural bioinformatics. In this research work, we introduced a new ab initio protein structure prediction approach based on the generalized pattern search approach that has been proved to be effective in many academic and real-world applications. We modeled the peptide structure prediction as a nonlinear optimization problem using the Gps algorithm to minimize a potential energy function, the Ecepp/3 function, to find the near-native structure of this kind of molecule, according to the thermodynamical hypothesis.

TABLE 4 Comparison between pattern-search based algorithms on the melittin peptide (PDB Id. 2MLT)

Algorithm	Secondary structures	Potential energy (kcal/mol)	RMSD C_α Å
Gps	yes	−104.349	3.089
MADS	yes	−94.780	3.378
Gps	no	−80.817	5.800
MADS	no	−78.124	6.050
PPSwarm	yes	100.312	7.431
PPSwarm	no	100.312	7.431

For each algorithm we report the best solution in terms of potential energy, and its relative RMSD C_α. PPswarm only returns conformations with positive energy values.

TABLE 5 2MLT

Klepeis et al. (1)						Gps					
Res	ϕ	ψ	Res	ϕ	ψ	Res	ϕ	ψ	Res	ϕ	ψ
1	69	−96	11	−74	−43	1	180	98	11	−134	72
2	−82	−28	12	−76	−30	2	−63	74	12	−101	−95
3	−66	−27	13	−148	78	3	−63	−40	13	−63	−15
4	−69	−27	14	−69	86	4	−67	−38	14	−73	−56
5	−83	−45	15	−154	173	5	−63	−46	15	−67	−28
6	−83	72	16	−57	−31	6	−67	−35	16	−68	−38
7	−64	−40	17	−56	−45	7	−65	−41	17	−65	−41
8	−66	−41	18	−82	−32	8	−63	−44	18	−68	−37
9	−70	−36	10	−68	−33	9	−64	−39	19	−60	−47
10	−76	−28	20	−79	−46	10	−78	−46	20	−76	−41

On the left, the dihedral angles of the configuration found by Klepeis et al. (1); and on the right, the one found by Gps.

TABLE 6 Testbed 1(9aa-13aa)

PDB Id.	Length	Energy (kcal/mol)	van der Waals (kcal/mol)	Solvation (kcal/mol)	RMSD$_{C\alpha}$ Å
1egs	9	−46.64	−28.81	−25.30	2.343
1c98	10	−82.40	−46.30	−37.71	3.925
1i83	11	−111.26	−51.80	−60.41	3.978
1i93	11	−134.00	−50.59	−79.21	3.858
1i98	11	−150.67	−54.82	−85.24	4.177
1qs3	11	−113.98	−51.53	−60.60	3.277
1qcm	11	−55.64	−23.63	−36.86	0.720
1m02	12	−118.26	−54.42	−64.08	3.055
1in3	12	−132.88	−60.15	−66.92	3.302
1cnl	12	−127.01	−58.75	−67.62	4.261
1l3q	12	−112.30	−54.51	−52.85	4.635
1d6x	13	−170.07	−78.88	−81.10	3.931
1g89	13	−142.77	−73.163	−66.82	2.501
1hje	13	−90.99	−50.39	−45.28	4.431
1im7	13	−78.73	−51.68	−39.96	4.354
1lcx	13	−160.05	−90.26	−53.26	2.547
1not	13	−123.78	−57.18	−67.22	2.973
1qfa	13	−165.03	−61.62	−95.77	4.295

Results obtained by Gps on the PEPstr benchmark: for each instance, we report the potential energy, the van der Waals term, and the Solvation energy term; moreover, we report the RMSD C_α for the best-found conformation in terms of potential energy.

The algorithm was tested on the same set of peptides used for the validation of the state-of-the-art algorithm known as PEPstr; the experiments show that Gps clearly outperforms PEPstr by 21.17% in terms of average RMSD C_α, and this result confirms that it is a suitable algorithm for the prediction of spatial conformations of bioactive molecules.

As future work we are currently investigating several research fronts. The first one is the understanding of how the bound settings may affect the Gps algorithm performance; in particular, whether the use of β-turn information can help Gps in finding good quality structures. Subsequently, we want to extend the algorithm by using the cluster analysis as a post-processing procedure to overcome some limitations of the algorithm; in particular, after the optimization process, we can apply this analysis to return the representative conformation of each cluster rather than just the conformation with the lowest potential energy value. This approach gives a human expert the chance to decide which one is the most biologically plausible conformation. Another research front is the use of a more powerful heuristic search procedure than the Latin hypercube sampling; the main exploring ability of the algorithm relies on the Search procedure, since the Poll phase acts as a local optimizer. It is obvious that introducing an algorithm with a good exploring ability is a crucial point toward improving the effectiveness of the algorithm. Finally, we want to smooth the potential energy function landscape by using a surrogate approach. We are working on a surrogate definition of the PSP, where we tackle the optimization of the surrogate function by using some derivative-based optimization tools or some quadratic programming procedures, and then by moving the solution to a nearby mesh point, hopefully to obtain a better next iterate. This is the approach used

TABLE 7 Testbed 2 (14aa-17aa)

PDB Id.	Length	Energy (kcal/mol)	van der Waals (kcal/mol)	Solvation (kcal/mol)	RMSD$_{C\alpha}$ Å
1a13	14	−81.31	−54.59	−42.30	4.480
1gjf	14	−134.76	−64.55	−71.06	4.800
1d7n	14	−82.85	−51.55	−26.33	1.883
1niz	14	−126.73	−54.39	−71.53	4.983
1dn3	15	−148.26	−75.05	−59.58	1.266
1gje	15	−142.22	−81.62	−60.48	3.784
2bta	15	−186.65	−70.24	−100.32	4.595
1akg	16	−114.01	−60.26	−54.13	3.872
1id6	16	−150.92	−73.69	−74.42	4.952
2bp4	16	−229.66	−108.58	−93.42	0.917
1e0q	17	−112.63	−57.65	−59.05	4.256

Results obtained by Gps on the PEPstr benchmark: for each instance, we report the potential energy, the van der Waals term, and the Solvation energy term; moreover, we report the RMSD C_α for the best-found conformation in terms of potential energy.

TABLE 8 Testbed 3 (18aa-20aa)

PDB Id.	Length	Energy (kcal/mol)	van der Waals (kcal/mol)	Solvation (kcal/mol)	RMSD$_{C\alpha}$ Å
1b03	18	−151.19	−48.43	−99.56	3.802
1d9m	18	−186.78	−119.02	−61.96	3.471
1hu5	18	−185.00	−78.53	−88.55	2.973
1pef	18	−199.27	−126.92	−51.64	0.595
1rpv	18	−325.44	−104.75	−169.72	1.988
1ien	19	−222.49	−115.12	−85.01	4.588
1jav	19	−222.99	−134.39	−72.91	2.082
1kzv	19	−192.77	−104.29	−71.12	2.087
1p0j	19	−207.27	−109.56	−84.33	2.538
1p0l	19	−217.24	−116.91	−85.23	1.497
1p5k	19	−238.50	−118.71	−93.09	1.581
1d9p	20	−188.79	−123.90	−55.44	2.281
1odp	20	−280.39	−133.94	−110.42	1.712
1sol	20	−204.45	−80.73	−88.919	3.060

Results obtained by Gps on the PEPstr benchmark: for each instance, we report the potential energy, the van der Waals term, and the Solvation energy term; moreover, we report the RMSD C_α for the best-found conformation in terms of potential energy.

TABLE 9 Comparison of PEPstr and Gps results

Algorithm	Model	Average RMSD C_α
PEPstr	I	7.1 Å
PEPstr	II	4.4 Å
PEPstr	III	4.1 Å
PEPstr	IV	4.0 Å
Gps	—	3.153 Å

PEPstr produces four models: model I uses extended conformations; model II uses regular secondary states; model III regular states and β-turns; and model IV extends model III using the χ_i angles from rotamer library (34). Gps outputs one model, and it uses secondary structure information only for instances with at least 15 amino acids. It turns out that Gps outperforms PEPstr of 21.17% on average RMSD C_α.

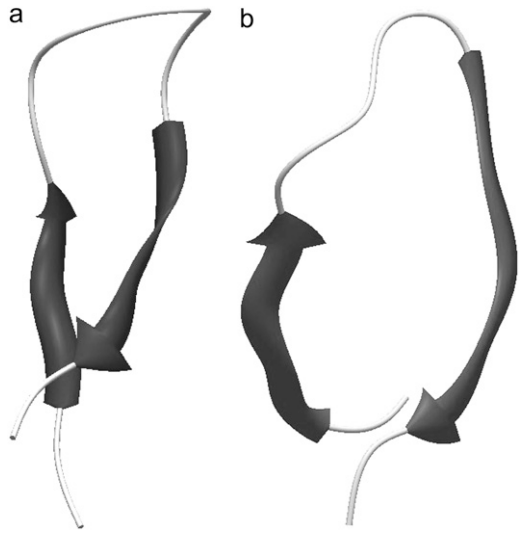

FIGURE 5 A comparison of the structure stored in the PDB (*a*) for the antibody-bound HIV-1IIIB peptide (PDB Id.: 1B03) and the structure predicted by Gps (RMSD $C_\alpha = 3.8021$Å) (*b*).

in the Boeing Design Explorer software (21), and it is a visionary research topic for the protein structure prediction problem.

REFERENCES

1. Bernal, J., and D. Crowfoot. 1934. X-ray photographs of crystalline pepsin. *Nature.* 133:794–795.
2. Wütrich, K. 1986. NMR of Proteins and Nucleic Acids. J. Wiley, New York.
3. Anfinsen, C. B., E. Haber, M. Sela, and F. H. White. 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA.* 47:1309–1314.
4. Levinthal, C. 1968. Are there pathways for protein folding? *J. Chim. Phys.* 65:44–45.
5. Shi, J., T. Blundell, and K. Mizuguchi. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* 310:243–257.
6. de Bakker, P., A. Bateman, D. Burke, R. Miguel, K. Mizuguchi, J. Shi, H. Shirai, and T. Blundell. 2001. HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics.* 17:748–749.
7. Lathrop, R. 1994. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* 7:1059–1068.
8. Rost, B., R. Schneider, and C. Sander. 1997. Protein fold recognition by prediction-based threading. *J. Mol. Biol.* 270:471–480.
9. Klepeis, J., M. Pieja, and C. Floudas. 2003. Hybrid global optimization algorithms for protein structure prediction: alternating hybrids. *Biophys. J.* 84:869–882.
10. Lee, J., H. Scheraga, and S. Rackovsky. 1997. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. *J. Comput. Chem.* 18:1222–1232.
11. Carloni, P., U. Rothlisberger, and M. Parrinello. 2002. The role and perspective of ab initio molecular dynamics in the study of biological systems. *Acc. Chem. Res.* 35:455–464.
12. Kihara, D., H. Lu, A. Kolinski, and J. Skolnick. 2001. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci. USA.* 98:10125–10130.
13. Cutello, V., G. Narzisi, and G. Nicosia. 2006. A multi-objective evolutionary approach to the protein structure prediction problem. *J. R. Soc. Interface.* 3:139–151.
14. Kim, D. E., D. Chivian, and D. Baker. 2004. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* 32:526–531.
15. Floudas, C. 2007. Computational methods in protein structure prediction. *Biotechnol. Bioeng.* 97:207–213.
16. Klepeis, J., and C. Floudas. 2003. ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.* 85:2119–2146.
17. Cutello, V., G. Morelli, G. Nicosia, and M. Pavone. 2005. Immune algorithms with aging operators for the string folding problem and the protein folding problem. 5th European Conference on Computation in Combinatorial Optimization (EvoCOP). 3448:80–90.
18. Audet, C., and J. E. Dennis, Jr. 2003. Analysis of generalized pattern searches. *SIAM J. Optim.* 13:889–903.
19. Torczon, V. 1997. On the convergence of pattern search algorithms. *SIAM J. Optim.* 7:1–25.
20. Lewis, R. M., and V. Torczon. 2000. Pattern search algorithms for linearly constrained minimization. *SIAM J. Optim.* 10:917–941.
21. Santner, T., B. Williams, and W. Notz. 2003. The Design and Analysis of Computer Experiments. Springer, New York.
22. Allison, J., B. Roth, M. Kokkolaras, I. Kroo, and P. Papalambros. 2006. Aircraft family design using decomposition-based methods. Proceedings of the 11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, September 6–8, 2006, Portsmouth, VA.
23. Abramson, M. 2004. Mixed variable optimization of a load-bearing thermal insulation system using a filter pattern search algorithm. *Optim. Eng.* 5:157–177.
24. Nemethy, G., K. Gibson, K. Palmer, C. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. Scheraga. 1992. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.* 96:6472–6484.
25. Zhang, Y. 2007. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins Struct. Funct. Bioinform.* 69:108–117.
26. Wu, S., and Y. Zhang. 2007. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* 35:33–75.
27. Zhang, Y., and J. Skolnick. 2004. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA.* 101:7594–7599.
28. Zhou, H., and Y. Zhou. 2004. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins Struct. Funct. Bioinform.* 55:1005–1013.
29. Zhou, H., and Y. Zhou. 2005. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins Struct. Funct. Bioinform.* 58:321–328.
30. Karplus, K., R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans, and R. Hughey. 2003. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins Struct. Funct. Bioinform.* 53:491–496.
31. Söding, J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics.* 21:951–960.
32. Chivian, D., and D. Baker. 2006. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Res.* 34:e112.
33. Simons, K., C. Kooperberg, E. Huang, and D. Baker. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268:209–225.
34. Dunbrack, R., Jr. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* 6:1661–1681.

35. Xu, J., M. Li, D. Kim, and Y. Xu. 2003. RAPTOR: optimal protein threading by linear programming. *J. Bioinform. Comput. Biol.* 1:95–117.

36. Kaur, H., A. Garg, and G. Raghava. 2007. PEPstr: a de novo method for tertiary structure prediction of small bioactive peptides. *Protein Pept. Lett.* 14:626–631.

37. Thomas, A., S. Deshayes, M. Decaffmeyer, M. Van Eyck, B. Charloteaux, and R. Brasseur. 2006. Prediction of peptide structure: how far are we? *Proteins Struct. Funct. Bioinform.* 65:889–897.

38. Case, D., D. Pearlman, J. Caldwell, T. Cheatham III, W. Ross, C. Simmerling, T. Darden, K. Merz, R. Stanton, A. Cheng, J. Vincent, M. Crowley, V. Tsui, R. Radmer, Y. Duan, J. Pitera, I. Massova, G. Seibel, U. Singh, P. Weiner, and P. Kollman. 1999. AMBER 6. University of California at San Francisco.

39. Kolda, T., R. Lewis, and V. Torczon. 2004. Optimization by direct search: new perspectives on some classical and modern methods. *SIAM Rev.* 45:385–482.

40. Booker, A., J. Dennis, P. Frank, D. Serafini, V. Torczon, and M. Trosset. 1999. A rigorous framework for optimization of expensive functions by surrogates. *Struct. Multidisc. Optimiz.* 17:1–13.

41. Audet, C., A. Booker, J. Dennis, Jr., P. Frank, and D. Moore. 2000. A surrogate-model-based method for constrained optimization. Proceedings of the Symposium on Multidisciplinary Analysis and Optimization. AIAA Paper.

42. Zhao, Z., J. Meza, and M. Van Hove. 2006. Using pattern search methods for surface structure determination of nanomaterials. *J. Phys. Condens. Matter.* 18:8693–8706.

43. Reference deleted in proof.

44. Pollastri, G., D. Przybylski, B. Rost, and P. Baldi. 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins Struct. Funct. Bioinform.* 47:228–235.

45. Lesk, A. 2001. Introduction to Protein Architecture: The Structural Biology of Proteins. Oxford University Press, New York.

46. Cornell, W., P. Cieplak, C. Bayly, I. Gould, K. Merz, D. Ferguson, D. Spellmeyer, T. Fox, J. Caldwell, and P. Kollman. 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117:5179–5197.

47. Hermans, J., H. Berendsen, W. Van Gusteren, and J. Postma. 1984. A consistent empirical potential for water-protein interactions. *Biopolymers.* 23:1513–1518.

48. Momany, F. A., R. F. McGuire, A. W. Burgess, and H. A. Scheraga. 1975. Energy parameters in polypeptides. VII Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions and intrinsic torsional potentials for the naturally occurring amino acids. *J. Phys. Chem.* 79:2361–2381.

49. Roterman, I., M. Lambert, K. Gibson, and H. Scheraga. 1989. A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. ϕ-ψ maps for *n*-acetyl alanine *n*′-methyl amide: comparisons, contrasts and simple experimental tests. *J. Biomol. Struct. Dyn.* 7:421–453.

50. Vila, J., R. Williams, M. Vasquez, and H. Scheraga. 1991. Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. *Proteins Struct. Funct. Bioinform.* 10:199–218.

51. Wesson, L., and D. Eisenberg. 1992. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.* 1:227–235.

52. Eisenberg, D., M. Wesson, and M. Yamashita. 1989. Interpretation of protein folding and binding with atomic solvation parameters. *Chem. Scr.* 29:217–221.

53. Juffer, A., F. Eisenhaber, S. Hubbard, D. Walther, and P. Argos. 1995. Comparison of atomic solvation parametric sets: applicability and limitations in protein folding and binding. *Protein Sci.* 4:2499–2509.

54. Schiffer, C., J. Caldwell, P. Kollman, and R. Stroud. 1993. Protein structure prediction with a combined solvation free energy-molecular mechanics force field. *Mol. Simul.* 10:121–149.

55. Eisenberg, D., and A. McLachlan. 1986. Solvation energy in protein folding and binding. *Nature.* 319:199–203.

56. Vin Freyberg, B., T. Richmond, and W. Breaun. 1993. Surface area included in energy refinement of proteins: a comparative study on atomic solvation parameters. *J. Mol. Biol.* 233:275–292.

57. Ooi, T., M. Oobatake, G. Nemethy, and H. Scheraga. 1987. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA.* 84:3086–3090.

58. Eisenmenger, F., U. Hansmann, S. Hayryan, and C. Hu. 2001. SMMP. A modern package for simulation of proteins. *Comput. Phys. Commun.* 138:192–212.

59. Stein, M. 1987. Large sample properties of simulations using Latin hypercube sampling. *Technometrics.* 29:143–151.

60. Audet, C., and J. Dennis, Jr. 2006. Mesh adaptive direct search algorithms for constrained optimization. *SIAM J. Optim.* 17:188–217.

61. Zemla, A., Č. Venclovas, J. Moult, and K. Fidelis. 1999. Processing and analysis of CASP 3 protein structure predictions. *Proteins Struct. Funct. Bioinform.* 37:22–29.

62. Siew, N., A. Elofsson, L. Rychlewski, and D. Fischer. 2000. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics.* 16:776–785.

63. Zhang, Y., and J. Skolnick. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinform.* 57:702–710.

64. Yona, G., and M. Levitt. 2000. A unified sequence-structure classification of protein sequences: combining sequence and structure in a map of the protein space. Proceedings of the Fourth Annual International Conference on Computational Molecular Biology.

65. Vaz, A., and L. Vicente. 2007. A particle swarm pattern search method for bound constrained global optimization. *J. Glob. Optim.* 39:197–219.

66. Marcote, I., F. Separovic, M. Auger, and S. Gagne. 2004. A multidimensional ^{1}H NMR investigation of the conformation of methionine-enkephalin in fast-tumbling bicelles. *Biophys. J.* 86:1587–1600.

67. Hansmann, U., and L. Wille. 2002. Global optimization by energy landscape paving. *Phys. Rev. Lett.* 88:68105.

68. Hough, P., T. Kolda, and V. Torczon. 2002. Asynchronous parallel pattern search for nonlinear optimization. *SIAM J. Sci. Comput.* 23: 134–156.

69. Zimmerman, S., M. Pottle, G. Némethy, and H. Scheraga. 1977. Conformational analysis of the 20 naturally occurring amino acid residues using ECEPP. *Macromolecules.* 10:1–9.

70. Lee, J., H. Scheraga, and S. Rackovsky. 1998. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. *Biopolymers.* 46:103–116.