JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2006) 3, 139–151
doi:10.1098/rsif.2005.0083
Published online 31 October 2005

A multi-objective evolutionary approach to the protein structure prediction problem

Vincenzo Cutello, Giuseppe Narzisi and Giuseppe Nicosia[†]

Department of Mathematics and Computer Science, University of Catania, V. le A. Doria 6, 95125 Catania, Italy

The protein structure prediction (PSP) problem is concerned with the prediction of the folded, native, tertiary structure of a protein given its sequence of amino acids. It is a challenging and computationally open problem, as proven by the numerous methodological attempts and the research effort applied to it in the last few years. The potential energy functions used in the literature to evaluate the conformation of a protein are based on the calculations of two different interaction energies: *local* (bond atoms) and *non-local* (non-bond atoms). In this paper, we show experimentally that those types of interactions are in conflict, and do so by using the potential energy function Chemistry at HARvard Macromolecular Mechanics. A multi-objective formulation of the PSP problem is introduced and its applicability studied. We use a multi-objective evolutionary algorithm as a search procedure for exploring the conformational space of the PSP problem.

Keywords: multi-objective optimization; Pareto front; protein folding; protein structure prediction; multi-objective evolutionary algorithms

1. INTRODUCTION

Proteins are long sequences of 20 different amino acids. Proteins are known to have many important functions in the cell, such as enzymatic activity, storage and transport of material, signal transduction, antibodies and more (Whisstock & Lesk 2003). The amino acids composition of a protein will usually uniquely determine its three-dimensional structure (Anfinsen 1973), to which the protein's functionality is directly related.

The protein structure prediction (PSP) problem is currently one of the most challenging problems in Biochemistry and Bioinformatics (Nicosia 2004). It is simply defined as the task of understanding and predicting how the information coded in the amino acid sequence of proteins translates into the three-dimensional structure of the biologically active protein. If we were able to solve the PSP problem we would greatly simplify, for example, the task of understanding the mechanism of hereditary and infectious diseases, of designing drugs with specific therapeutic properties and of growing biological polymers with specific material properties.

Protein folding (PF) has to be distinguished from PSP. In the PSP, we are not interested in the folding process (dynamic aspect), but only in the attained final structure (static aspect). Common methods for finding protein three-dimensional structures (such as X-ray crystallographic and NMR—nuclear magnetic resonance) are slow and costly, and may take up to several months of lab work. As a consequence, there has been a continuously growing interest in the design of *ad hoc* algorithms for the PSP problem. There are two main types of computational strategies which are employed today: knowledge-based and *ab initio*. The hypothesis behind knowledge-based methods (homology modelling, Tramontano & Morea 2003; threading, Mirny *et al.* 2000) is that similar sequences will fold in a similar way. *Ab initio* strategies (Klepeis *et al.* 2005) are required when no homology is available so that one is forced to fold the proteins from scratch.

Successful structure prediction requires a free energy function sufficiently close to the right potential for the native state, as well as a method for exploring conformational space. Current potential functions, however, have limited accuracy and the conformational space is vast. Several algorithmic approaches have been applied to the PSP problem in the last 20 years (Nicosia 2004; Cozzetto *et al.* 2005): molecular dynamics, Metropolis Monte Carlo, simulated annealing, simulated tempering, evolutionary algorithms. In spite of all these efforts, PSP remains a challenging computationally open problem.

2. MULTI-OBJECTIVE OPTIMIZATION

Historically, PF and PSP, both central problems in molecular biology, have been approached as a large single-objective optimization problem: given the primary sequence find the three-dimensional native conformation with minimum energy (PSP), and the pathways to reach the native conformation (PF), using a single-objective potential energy function. Molecular dynamics, Monte Carlo methods and evolutionary

[†] Author for correspondence (nicosia@dmi.unict.it).

One contribution of 8 to a themed supplement 'Statistical mechanics of molecular and cellular biological systems'.

Received 4 May 2005
Accepted 22 August 2005

algorithms (Bowie & Eisemberg 1994; Hansmann & Okamoto 1997; Pendersen & Moult 1997; Simons *et al.* 1997; Cui *et al.* 1998; Nicosia 2004) are today's state of the art methodologies to tackle PF and PSP as a single-objective optimization problem.

We conjecture and partially verify by computational experiments that it could be more suitable to model the PSP problem as a multi-objective optimization problem (MOOP).

When an optimization problem involves more than one objective function, the task of finding one (or more) optimum solution, is known as multi-objective optimization (Steuer 1986). The PSP problem naturally involves multiple objectives. Different solutions, i.e. the three-dimensional conformations, may involve a trade-off (the conflicting scenario in the funnel landscape) among different objectives. An optimum solution with respect to one objective may not be optimum with respect to another objective. Consequently, one cannot choose a solution which is optimal with respect to only one objective. In general, in problems with more than one conflicting objective, there is no single optimum solution. There exists, instead, a set of solutions which are all optimal, called the *optimal Pareto front*.

Hence, for a MOOP we can define the following procedure:

(i) find the optimal (or the observed) Pareto front with a wide range of values for objectives; and
(ii) choose one of the solutions in the Pareto front, using some higher-level information.

We would like to emphasize the fact that our major goal in using the multi-objective approach to PSP is to find the *folded ensemble* by means of the Pareto optimality concept.

We think that finding the native structure of a given protein is not equivalent 'to finding a native state needle in a conformational space haystack' but, instead, should be more like 'finding a set of equivalent needles in a haystack'. Obviously, the problem is still far from being solved, but we are modelling the PSP problem in an alternative and more accurate way; that is, at any stage the molecule exists in an ensemble of conformations. We want to model such a stage as an approximated Pareto front.

Indeed, the authors of the paper (Ma *et al.* 1999) report: 'The long-held views on lock-and-key versus induced fit in binding arose from the notion that a protein exists in a single, most stable conformation, dictated by its sequence. However, in solution proteins exist in a range of conformations, which may be described by statistical mechanical laws and their populations follow statistical distributions. Upon binding, the equilibrium will shift in favour of the bound conformation from the ensemble of conformations around the bottom of the folding funnel.'

Hence, the ensemble of equivalent conformations is crucial for proteins in solution, as well. Finally, it is evident how the multi-objective approach intends to discover the conformation populations around the bottom of the folding funnel using the Pareto optimality concept so as to study the biological activity in this stage. Indeed, conformational diversity around the bottom of the folding funnel may provide a simple yet elegant solution to a range of binding processes. We want to discover this conformational diversity around the bottom of the funnel using the Pareto optimality.

Since tackling multi-objective problems in their native form (i.e. rather than adding together objective values or treating some of them as constraints) can be actually beneficial, it should be more effective and accurate to face a problem as a MOOP when the objectives have a significant anti-correlation or conflict, although this is probably not a sufficient condition (Louis & Rawlins 1993; Knowles *et al.* 2001; Jensen 2003). In particular, Knowles *et al.* (2001) showed experimentally how in some cases transforming a single-objective problem into a multi-objective one can reduce the number of local optima and facilitate the optimization process. Moreover, they showed how this process, called *multi-objectivization*, can facilitate the search process for neighbourhood based algorithms, like Pareto archived evolutionary strategy (PAES).

In a recent article (Day *et al.* 2002), Lamont and co-authors reformulated the PSP problem as a MOOP and used a multi-objective evolutionary algorithm (MO fmGA) for the structure prediction of two small protein sequences: [Met]-enkephelin (five residues), polyalanine (14 residues). After this initial approach, this idea was applied to medium size protein sequences (46–70 residues) with promising results (see Cutello *et al.* 2005).

In the following, we formally introduce the MOOP.

2.1. Multi-objective optimization problems

A MOOP can be formally defined as follows.

Definition 2.1. *Find a vector* $\boldsymbol{x}^* = [x_1^*, x_2^*, \ldots, x_n^*]^{\mathrm{T}}$ *which*

(i) *satisfies the p equality constraints,*

$$h_i(\boldsymbol{x}) = 0, \quad i = 1, 2, \ldots, p; \tag{2.1}$$

(ii) *is subject to the m inequality constraints,*

$$g_i(\boldsymbol{x}) \geq 0, \quad i = 1, 2, \ldots, m; \tag{2.2}$$

(iii) *and which optimizes the vector function,*

$$\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x})]^{\mathrm{T}}. \tag{2.3}$$

Hence, we have the following two hyperspaces.

Definition 2.2. *Equations (2.1) and (2.2) define the feasible region (or decision variable space)*

$$\begin{aligned} \varOmega = \{\boldsymbol{x} \in \Re^n : g_i(\boldsymbol{x}) \geq 0, \quad i = 1, 2, \ldots, m; \\ h_i(\boldsymbol{x}) = 0, \quad i = 1, 2, \ldots, p\}, \end{aligned} \tag{2.4}$$

and any point $\boldsymbol{x} \in \varOmega$ *defines a feasible solution.*

Definition 2.3. *The vector function* $\boldsymbol{f}(\boldsymbol{x})$ *maps the elements of* $\varOmega$ *into a set* $\varLambda$ *which represents all possible values of the objective functions:*

$$\varLambda = (\boldsymbol{f}(\boldsymbol{x}) \in \Re^k : \boldsymbol{x} \in \varOmega). \tag{2.5}$$

The evaluation function of the MOOP $f: \Omega \to \Lambda$, maps decision variables $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ to vectors $\boldsymbol{y}=(y_1, y_2, \ldots, y_k)$.

Definition 2.4. *A point* $\boldsymbol{x}^* \in \Omega$ *is Pareto optimal if for every* $\boldsymbol{x} \in \Omega$ *and* $I=\{1, 2, \ldots, k\}$ *either*

$$\forall_{i \in I}(f_i(\boldsymbol{x}) = f_i(\boldsymbol{x}^*)) \tag{2.6}$$

or there is at least one $i \in I$ *such that*

$$f_i(\boldsymbol{x}) \geq f_i(\boldsymbol{x}^*). \tag{2.7}$$

Definition 2.5. *A vector* $\boldsymbol{u}=(u_1, \ldots, u_k)$ *is said to dominate* $\boldsymbol{v}=(v_1, \ldots, v_k)$*, denoted by*

$$\boldsymbol{u} \preccurlyeq \boldsymbol{v},$$

if and only if $\boldsymbol{u}$ *is partially less than* $\boldsymbol{v}$*, i.e. for all* $i \in \{1, \ldots, k\}$, $u_i \leq v_i \wedge \exists i \in \{1, \ldots, k\} : u_i < v_i$.

If the vector $\boldsymbol{u}$ dominates the vector $\boldsymbol{v}$, or mathematically $\boldsymbol{u} \preccurlyeq \boldsymbol{v}$, we also say that $\boldsymbol{v}$ is dominated by $\boldsymbol{u}$, or $\boldsymbol{u}$ is non-dominated by $\boldsymbol{v}$, or $\boldsymbol{u}$ is not inferior to $\boldsymbol{v}$.

Definition 2.6. *A point* $\boldsymbol{x}^* \in \Omega$ *is a weakly non-dominated solution if there is no* $\boldsymbol{x} \in \Omega$ *such that* $f_i(\boldsymbol{x}) < f_i(\boldsymbol{x}^*)$*, for* $i=1, \ldots, k$.

Definition 2.7. *A point* $\boldsymbol{x}^* \in \Omega$ *is a strongly non-dominated solution if there is no* $\boldsymbol{x} \in \Omega$ *such that* $f_i(\boldsymbol{x}) \leq f_i(\boldsymbol{x}^*)$*, for* $i=1, \ldots, k$ *and for at least one value of* i, $f_i(\boldsymbol{x}) < f_i(\boldsymbol{x}^*)$.

Thus, if $\boldsymbol{x}^*$ is strongly non-dominated, it is also weakly non-dominated, but the converse is not necessarily true.

Definition 2.8. *For a given MOOP* $\boldsymbol{f}(\boldsymbol{x})$*, the Pareto optimal set,* $\mathcal{P}^*$*, is defined as*

$$\mathcal{P}^* = \{\boldsymbol{x} \in \Omega : \neg \exists\, \boldsymbol{x}' \in \Omega \quad \boldsymbol{f}(\boldsymbol{x}') \preccurlyeq \boldsymbol{f}(\boldsymbol{x})\}. \tag{2.8}$$

In this paper, a Pareto optimal set that truly meets this definition is called a true Pareto optimal set, $\mathcal{P}^*_{\text{true}}$. In contrast, a Pareto optimal set that is obtained by means of an optimization method is referred to as an observed Pareto optimal set, $\mathcal{P}^*_{\text{obs}}$. In reality, an observed Pareto optimal set is an *estimate* (or a discrete representation) of a true Pareto optimal set.

Identifying a good estimate $\mathcal{P}^*_{\text{obs}}$ is the key factor for the decision-maker's selection of a compromise solution, which satisfies the objectives as much as possible.

We denote the observed Pareto optimal set at time-step t obtained using an optimization method by $\mathcal{P}^{*,t}_{\text{obs}}$ (or the current observed Pareto optimal set). Moreover, we have

$$\mathcal{P}^{*,t}_{\text{obs}} = \{\boldsymbol{x}^t_1, \ldots, \boldsymbol{x}^t_{np}\}, \tag{2.9}$$

where $np = |\mathcal{P}^{*,t}_{\text{obs}}|$ is the total number of observed Pareto solutions at time-step t.

Obviously, the major problem a decision-maker needs to solve, is to find the best

$$\boldsymbol{x} \in \mathcal{P}^*_{\text{obs}}.$$

Definition 2.9. *For a given MOOP* $\boldsymbol{f}(\boldsymbol{x})$ *and Pareto optimal set* $\mathcal{P}^*$*, the Pareto front,* $\mathcal{PF}^*$*, is defined as*

$$\mathcal{PF}^* = \{\boldsymbol{u} = \boldsymbol{f} = (f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x})) | \boldsymbol{x} \in \mathcal{P}^*\}. \tag{2.10}$$

As for the Pareto optimal set, we can define the *observed Pareto front* at time-step t by an optimization method:

$$\mathcal{PF}^{*,t}_{\text{obs}} = \{\boldsymbol{u}^t_1, \boldsymbol{u}^t_2, \ldots, \boldsymbol{u}^t_N\}, \tag{2.11}$$

where $N = |\mathcal{PF}^{*,t}_{\text{obs}}|$ is the total number of observed Pareto front solutions at time-step t.

The goal of our work is to estimate the observed Pareto front, $\mathcal{PF}^{*,t}_{\text{obs}}$, by a multi-objective evolutionary algorithm for the structure prediction of real proteins. Identifying a good estimate of $\mathcal{PF}^{*,t}_{\text{obs}}$ is crucial for the biologist's selection of a stable fold protein near native conformation, under biological conditions, satisfying the objectives as much as possible.

3. METHODS

The most difficult task when using a search procedure for the PSP problem is to come up with good:

(i) *representation* of the conformations,
(ii) *cost function* for evaluating conformations, and
(iii) *metrics* to evaluate how similar to the native structure are the predicted conformations.

We will now introduce the problems correlated with those aspects and we will describe our choices.

3.1. Representation of the polypeptide chain

Few conformation-representations are commonly used:

(i) all-atom three-dimensional coordinates;
(ii) all-heavy-atom coordinates;
(iii) backbone atom coordinates + side-chain centroids;
(iv) C_α coordinates; and
(v) backbone and side-chain torsion angles.

Some algorithms use multiple representations and move among them for different purposes.

In this work, we use an internal coordinates representation (torsion angles), based on the fact that 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. All the ω torsion angles are fixed at their ideal value 180°. So, the degrees of freedom in this representation are the backbone and side-chain torsion angles (ϕ, ψ and χ_i). The number of χ angles depends on the residue type (see table 1).

3.2. Potential energy function

In order to evaluate the structure of a molecule we need to use some cost or energy functions. To come up with some good functions it would be natural to use quantum mechanics, but it is too computationally complex to be practical in modelling larger systems, so, we use classical physics. Sometimes called potential energy functions or force fields, these functions return a value for the energy based on the conformation of the molecule. They provide information on what conformations of

Table 1. Number of χ_i angles required to fix the positions of side-chain atoms in each residue type.

residue	number of χ angles
GLY, ALA, PRO	main chain
SER, CYS, THR, VAL	χ_1
ILE, LEU, ASP, ASN, HIS, PHE, TYR, TRP	χ_1, χ_2
MET, GLU, GLN	χ_1, χ_2, χ_3
LYS, ARG	χ_1, χ_2, χ_3, χ_4

the molecule are better or worse. The lower the energy value, then the better should be the conformation.

Most typical energy functions have the form

$$E(R) = \sum_{\text{bonds}} B(\boldsymbol{R}) + \sum_{\text{angles}} A(\boldsymbol{R}) + \sum_{\text{torsions}} T(\boldsymbol{R}) + \sum_{\text{non-bonded}} N(\boldsymbol{R}), \tag{3.1}$$

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

The literature on cost functions is enormous (Momany *et al.* 1975; Hermans *et al.* 1984; Cornell *et al.* 1995). In this work, in order to evaluate the conformation of a protein, we use the Chemistry at HARvard Macromolecular Mechanics (CHARMM) (v.27) energy function. CHARMM is a popular all-atom force field used mainly for studying macromolecules (MacKerell *et al.* 1998; Foloppe & MacKerell 2000). It is a composite sum of several molecular mechanics equations grouped into two major types: *bonded* (stretching, bending, torsion, Urey-Bradley, impropers) and *non-bonded* (van-der-Walls, electrostatics).

The CHARMM energy function has the form

$$\begin{aligned} E_{\text{charmm}} = {} & \underbrace{\sum_{\text{bonds}} k_{\text{b}}(b-b_0)^2}_{E_1} + \underbrace{\sum_{\text{UB}} k_{\text{UB}}(S-S_0)^2}_{E_2} \\ & + \underbrace{\sum_{\text{angles}} k_\theta(\theta-\theta_0)^2}_{E_3} + \underbrace{\sum_{\text{torsions}} k_\chi[1+\cos(n\chi-\delta)]}_{E_4} \\ & + \underbrace{\sum_{\text{impropers}} k_{\text{imp}}(\phi-\phi_0)^2}_{E_5} \\ & + \underbrace{\sum_{\text{non-bond}} \varepsilon_{ij}\left[\left(\frac{R\min_{ij}}{r_{ij}}\right)^{12} - \left(\frac{R\min_{ij}}{r_{ij}}\right)^{6}\right]}_{E_6} \\ & + \underbrace{\frac{q_i q_j}{e r_{ij}}}_{E_7}, \end{aligned} \tag{3.2}$$

where

(i) b is the bond length, b_0 is the bond equilibrium distance and k_{b} is the bond force constant;

(ii) S is the distance between two atoms separated by two covalent bonds (1, 3 distance), S_0 is the equilibrium distance and k_{UB} is the Urey Bradley force constant;

(iii) θ is the valence angle, θ_0 is the equilibrium angle and K_θ is the valence angle force constant;

(iv) χ is the dihedral or torsion angle, k_χ is the dihedral force constant, n is the multiplicity and δ is the phase angle;

(v) ϕ is the improper angle, ϕ_0 is the equilibrium improper angle and k_{imp} is the improper force constant; and

(vi) ε_{ij} is the Lennard Jones well depth, r_{ij} is the distance between atoms i and j, $R\min_{ij}$ is the minimum interaction radius, q_i is the partial atomic charges and e is the dielectric constant.

Typically, ε_i and $R\min_i$ are obtained for individual atom types and then combined to yield ε_{ij} and $R\min_{ij}$ for the interacting atoms via combining rules. In CHARMM, ε_{ij} values are obtained via the geometric mean $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$, and $R\min_{ij}$ via the arithmetic mean, $R\min_{ij} = (R\min_i + R\min_j)/2$.

As a final note, we would like to underline the fact that the CHARMM energy function simply adds together bond and non-bond energies. As it will be experimentally shown later on, this produces an energy landscape which does not well correlate with the real molecule folding process.

3.3. Distance matrix error and root mean square deviation metrics

To evaluate how similar the predicted conformation is to the native one, we employ root mean square deviation (RMSD) coupled with another frequently used metric, the distance matrix error (DME). RMSD is given by the formula

$$\text{RMSD}(a, b) = \sqrt{\frac{\sum_{i=1}^{n} |r_{ai} - r_{bi}|^2}{n}}, \tag{3.3}$$

where r_{ai} and r_{bi} are the positions of atom i of structure a and structure b, respectively, and where structures a and b have been optimally superimposed. Fitting was performed using the McLachlan algorithm (McLachlan 1982).

DME is given by the formula

$$\text{DME}(a, b) = \frac{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} (|r_{ai} - r_{aj}| - |r_{bi} - r_{bj}|)^2}}{n}. \tag{3.4}$$

This calculation does not require the superposition of coordinates. RMSD, which measures the similarity of atomic positions, is usually larger than DME, which measures the similarity of inter-atomic distances.

RMSD is one of the most used instruments for structure comparison. However, using RMSD alone has some negative aspects: best alignment does not always mean minimal RMSD; significance of RMSD depends on the size of the structures; significance of RMSD varies with protein type; it is not a good measure when all equivalent parts of the proteins cannot be simultaneously superposed; all atoms are usually treated equally, though, for example, residues on the surface have a higher degree of freedom than those in the core (so weights might be used in the calculation).

4. THE MULTI-OBJECTIVE FORMULATION

In order to reduce the size of the conformational space, backbone torsion angles are bounded in regions derived from secondary and supersecondary structure prediction (table 2).

Supersecondary structure is defined as the combination of two secondary structural elements with a short connecting peptide between one to five residues in length. A short connecting peptide can have a large number of conformations. They play an important role in defining protein structures. The conformations of the residues in the short connecting peptides are classified into five major types, namely, *a*, *b*, *e*, *l* or *t* (Sun & Jang 1996) each represented by a region on the ϕ–ψ map. Sun *et al.* (1997) developed an artificial neural network method (ANN) to predict the 11 most frequently occurring supersecondary structures: H–*b*–H, H–*t*–H, H–*bb*–H, H–*ll*–E, E–*aa*–E, E–*ea*–E, H–*lbb*–H, H–*lba*–E, E–*aal*–E, E–*aaal*–E and H–*l*–E, where H and E represent α-helix and β-strand, respectively.

Side-chain torsion angles are constrained in regions derived from the backbone-independent rotamer library of Roland L. Dunbrack (Dunbrack & Cohen 1997). Side-chain constraint regions are of the form: $[m-\sigma, m+\sigma]$; where m and σ are the mean and the s.d. for each side-chain torsion angle computed from the rotamer library. Under these constraints, the conformation is still highly flexible and the structure can take on various shapes that are vastly different from the native shape.

We can think of a protein as a collection of atoms linked by a chemical bond. With the symbol $a_i \leftrightarrow a_j$ we represent a chemical bond between the two atoms a_i and a_j. Using this notation we can divide all the atoms into two categories: *bond atoms* and *non-bond atoms*,

$$A_{\text{bond}} = \{\langle a_i, \ldots, a_{i+k}\rangle \mid \exists\, a_i \leftrightarrow a_{i+1}\}, \quad \forall\, i = 1 \ldots k, \quad 1 \le k \le 3, \tag{4.1}$$

$$A_{\text{non-bond}} = \{\langle a_i, \ldots, a_j\rangle \mid \neg\exists a_i \leftrightarrow a_j\}, \quad \forall i, j. \tag{4.2}$$

The bond set A_{bond} represents the set of all atom chains of max length four, in this way we consider only bonds, angles and torsion interactions between atoms, *local interaction*. The $A_{\text{non-bond}}$ set represents all the atoms not connected by chemical bond, which are atoms separated by at least three or more covalent bonds, *non-local interaction*. This division reflects the decomposition of CHARMM in two partial sums: bonded and non-bonded atom energies, following definition (3.2),

$$f_1 = E_{\text{bond}}(A_{\text{bond}}, \boldsymbol{C}_{\text{bond}}) = \sum_{k=1}^{5} E_k, \tag{4.3}$$

$$f_2 = E_{\text{non-bond}}(A_{\text{non-bond}}, \boldsymbol{C}_{\text{non-bond}}) = \sum_{k=6}^{7} E_k, \tag{4.4}$$

where symbols $\boldsymbol{C}_{\text{bond}}$ and $\boldsymbol{C}_{\text{non-bond}}$ are, respectively, the force constants involved for bond and non-bond atoms in equation (3.2).

Table 2. Corresponding regions of the secondary and supersecondary structure constraints.

supersecondary structures	ϕ	ψ
H (α-helix)	[−75°, −55°]	[−50°, −30°]
E (β-strand)	[−130°, −110°]	[110°, 130°]
a	[−150°, −30°]	[−100°, 50°]
b	[−230°, −30°]	[100°, 200°]
e	[30°, 130°]	[130°, 260°]
l	[30°, 150°]	[−60°, 90°]
t	[−160°,−50°]	[50°, 100°]
undefined	[−180°, 0°]	[−180°, 180°]

The bond energy characterizes the interactions between residues that are neighbours along the primary sequence. The non-bond term represents the interaction between residues that are separated in the primary sequence by at least two intervening residues (one to four interactions).

These two functions represent our minimization *objectives*, the torsion angles of the protein are the *decision variables* of the multi-objective problem, and the constraint regions are the *variable bounds.*

5. THE MULTI-OBJECTIVE EVOLUTIONARY ALGORITHM

The algorithm PAES was proposed for the first time by Knowles & Corne (1999). PAES is a multi-objective optimizer which uses a simple (1+1) local search evolution strategy. Nonetheless, it is capable of finding diverse solutions in the Pareto optimal set because it maintains an archive of non-dominated solutions which it exploits to accurately estimate the quality of new candidate solutions. At any iteration t, a candidate solution c_t and a mutated solution m_t must be compared for dominance. Acceptance is simple if one solution dominates the other. If neither solution dominates the other, the new candidate solution is compared with the reference population of previously archived non-dominated solutions. If the comparison fails to favour one solution over the other, the chosen solution is the one which resides in the least crowded region of the space. A maximum size of the archive is always maintained. The crowding procedure is based on recursively dividing up the M-dimensional objective space in 2^d equal-sized hypercubes, where d is a user defined depth parameter. The algorithm continues until a given, fixed number of *iterations* is reached.

I-PAES is a modified version of PAES with a different solution representation (polypeptide chain) and immune inspired operators: *cloning* and *hypermutation* (Cutello & Nicosia 2004). The algorithm starts by initializing a random conformation. The torsion angles (ϕ, ψ, χ_i) are generated randomly from the constraint regions. After that, the energy of the conformation (a point in the landscape) is evaluated. First, the protein structure in internal coordinates (torsion angles) is transformed in Cartesian coordinates. Then the CHARMM energy potential of the structure is

```
I-PAES($depth, archive\_size, objectives$)
1. t := 0;
2. Initialize($c$); /*Generate initial random solution*/
3. Evaluate($c$); /*Evaluation of initial solution*/
4. AddToArchive($c$); /*Add $c$ to archive*/
5. while(not(Termination()))
       /*Start Immune phase*/
6.     ($c_1^{clo}, c_2^{clo}$) := Cloning($c$); /*Clonal expansion phase*/
7.     ($c_1^{hyp}, c_2^{hyp}$) := Hypermutation($c_1^{clo}, c_2^{clo}$); /*Affinity maturation phase*/
8.     Evaluate($c_1^{hyp}, c_2^{hyp}$); /*Evaluation phase*/
10.    if($c_1^{hyp}$ dominates $c_2^{hyp}$) $m := c_1^{hyp}$;
10.    else if($c_2^{hyp}$ dominates $c_1^{hyp}$) $m := c_2^{hyp}$;
10.    else $m$ := Best($c_1^{hyp}, c_2^{hyp}$); /*min $E_{charmm}$ selection*/
12.        AddToArchive(Worst($c_1^{hyp}, c_2^{hyp}$)); /*max $E_{charmm}$ selection*/
       /*End Immune phase*/
       /*Start (1+1)-PAES*/
10.    if($c$ dominates $m$) discard $m$;
11.    else if($m$ dominates $c$)
12.        AddToArchive($m$);
13.        $c := m$;
14.    else if($m$ is dominated by any member of the archive) discard $m$;
15.    else test($c, m, archive\_size, depth$);
16.    t := t + 1;
17. endwhile
```

Figure 1. Pseudo-code of I-PAES.

computed using routines from TINKER Molecular Modelling Package (http://dasher.wustl.edu/tinker/).

At this point, we have the main loop of the algorithm. From the current solution, two clones will be generated, producing the solutions (c_1^{clo}, c_2^{clo}) which will be mutated into (c_1^{hyp}, c_2^{hyp}). After evaluation, the best clone (min E_{charmm}) between c_1^{hyp} and c_2^{hyp} is selected as new mutated solution m, while the other one, if possible, is added to the archive following the standard method of PAES to update the archive. From this moment on, the algorithm proceeds following the standard structure of PAES. Figure 1 shows the pseudo-code of the algorithm.

Two kinds of mutation operators were used in the affinity maturation phase (line 6 of I-PAES). The first clone is mutated using the first mutation operator and the second clone using the second mutation operator. The first mutation operator, M_1, may change the conformation dramatically. When this operator acts on a peptide chain, all the values of the backbone and side-chain torsion angles of a randomly chosen residue are re-selected from their corresponding constrained regions. The probability for the application of this operator is regulated by the following law:

$$M_1(\text{ffe}) = \exp\left\{\frac{-2 \times (\text{ffe})}{T_{\max}}\right\}, \tag{5.1}$$

where $T_{\max}$ is the max number of evaluation allowed and ffe is the number of fitness function evaluation done. The probability of mutation decreases as the search method proceeds. The second mutation operator, M_2, performs a local search of the conformational space. It will perturb some torsion angles (ϕ, ψ, χ_i) of a randomly chosen residue with the law

$$\theta' = \theta + N(0,1), \tag{5.2}$$

where θ is the generic torsion angle, and $N(0, 1)$ is a real number generated by a Gaussian distribution of mean $\mu = 0$ and s.d. $\sigma = 1$. The mutation rate used is similar to the scheme presented in (Cui *et al.* 1998). The number of mutations decreases as the search method proceeds following the law

$$M_2(\text{ffe}) = 1 + \left(\frac{L}{k}\right)\exp\left\{\frac{-2 \times (\text{ffe})}{T_{\max}}\right\}, \tag{5.3}$$

where ffe and $T_{\max}$ are defined as before, L is the number of residues and k is a constant set to 4.

5.1. Bond energy versus non-bond energy

Before we analyse the quality of the obtained results, we would like to experimentally validate such a multi-objective approach. It is based on the fact that local interaction (bond energy) and non-local interaction (non-bond energy) among atoms are in conflict. This is the typical characteristic of a MOOP. The literature on energy functions and about those two different interactions is very vast. Most of the major energy functions are based on the combined usage of bond and non-bond energies. There is no *formal proof*, however, about the conflict between them. We start by describing the simple intuition about the conflict and then we show how it is possible to verify it experimentally.

In the PF process, it was demonstrated experimentally that the native structure of a protein is at its global minimum of the thermodynamical potential (free energy) of the protein (Anfinsen 1973). This is a valid principle that governs the protein conformational search. During the pathway to reach the native structure, the protein is forced to decide what to do next. It is quite clear that it is possible to make movements that locally are able to decrease the bond energy of the system. Globally, however, this could not

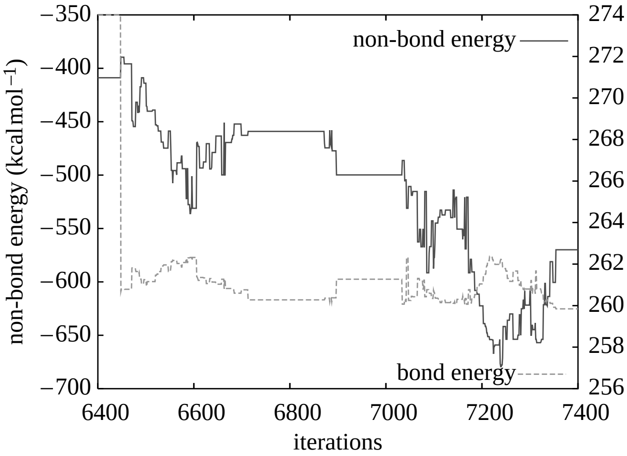

Figure 2. Conflict landscape between bond and non-band energies. Left y-axis range is for non-band energy, right y-axis range is for bond energy.

be true. For example, the electrostatic interactions or the short distance van der Waals interaction of atoms near in the space but far in primary sequence could be penalized although an improvement it is reach in bond, angles or torsion energies. This simple intuition is demonstrated experimentally by the plot in figure 2. In this figure, we show the typical bond and non-bond energy course during the iterations of the algorithm. It is clear that the two functions are in conflict. If one interaction decreases the other always increases but while this process goes on, there are compensation effects that will cause the minimization of the total energy.

5.2. Decision-making phase

As described in §2, after a Pareto front is found, to solve effectively a MOOP one has to choose a solution (or a class of solutions) in the Pareto front using some ‘higher-level information’. Such a *decision-making* phase, could be really difficult to accomplish, in particular when the number of objectives and solutions is large. Although, there is no universally accepted method, in general the most interesting solutions of the observed Pareto front are characterized by the fact that a small improvement in one objective will cause a large deterioration in at least one other objective. These solutions are called *knees* (Branke *et al.* 2004; Handl & Knowles 2005). In particular, in Branke *et al.* (2004), the authors describe an algorithm for finding the knees in the Pareto front: an angle-based method which uses the four closest neighbours. More in details, given a conformation point P, if we denote by A_1 and A_2 the two closest points from the left, and by B_1 and B_2 the two closest points from the right, we can form four angles: $\widehat{A_1PB_1}$, $\widehat{A_1PB_2}$, $\widehat{A_2PB_1}$ and $\widehat{A_2PB_2}$. The greatest of these four angles is then assigned to P. The knees of the Pareto front are the angles greater than a given threshold.

Using such a simple idea, we can adopt the following decision-making scheme:

(i) first, detect the solutions which lie in the knees of the observed Pareto front, using the angle-based method with four neighbours described above; and

(ii) select the solution with the lowest energy function value from these samples.

As we will see later, such a simple method is able to select solutions with a good trade-off between energy and metrics values (DME and RMSD).

This is just one possible approach for the decision-making phase. It is possible to use other type of higher-level information to select solutions from the Pareto front, using for instance structure stability, compactness, hydrophobic score, etc.

6. RESULTS

In this section, we report the results obtained using the multi-objective approach for PSP. We applied our algorithm to a famous short peptide ([Met]-enkephalin) and then to four proteins sequences from the Protein Data Bank (PDB). Table 7 shows the results for each protein. For 1ZDD, 1ROP, 1UTG and 1CRN proteins we set the maximum number of iterations to 2.5×10^5; while for Met-enkephalin peptide we ran the I-PAES for 3.5×10^5 energy functions evaluations.

[Met]-enkephalin. [Met]-enkephalin is a very short polypeptide, with only five amino acids (TYR–GLY–GLY–PHE–MET), 22 variable backbone and side-chain torsion (or dihedral) angles and 75 atoms. From an optimization point of view, the [Met]-enkephalin polypeptide is a paradigmatic example of multiple-minima problem. It is estimated to have more than 10^{11} locally optimal conformations. This peptide is an obvious ‘test bed’, for which a substantial amount of *in silico* experiments has been done (Li & Scheraga 1988). Figure 3 shows the dynamic of the Pareto fronts at different time-steps of the algorithms. Figure 4*a* instead shows the overlap between predicted and the Scheraga conformations (a classical benchmarks); while, figure 4*b* shows the overlap between predicted conformation and the native structure of the peptide 1PLW. After computing the Pareto front using our algorithm, firstly we detect the class of solutions in the knees of the observed Pareto front and then select the solution with lowest energy value. For the Met-enkephalin, the lowest energy value in the knees corresponds to the lowest energy value of the overall Pareto front, $-20.56\ \mathrm{kcal\ mol}^{-1}$; this conformation matches the Scheraga structure with $\mathrm{DME}_{\text{all-atoms}}=2.211$ Å, $\mathrm{RMSD}_{\text{all-atoms}}=2.83$ Å, $\mathrm{DME}_{C_\alpha}=0.454$ Å and $\mathrm{RMSD}_{C_\alpha}=0.490$ Å, and the crystal structure of 1PLW with $\mathrm{DME}_{\text{all-atoms}}=2.311$ Å, $\mathrm{RMSD}_{\text{all-atoms}}=3.605$ Å, $\mathrm{DME}_{C_\alpha}=1.200$ Å and $\mathrm{RMSD}_{C_\alpha}=1.740$ Å.

Disulphide-stabilized mini protein A domain (1ZDD). 1ZDD is a two-helix peptide of 34 residues (Starovasnik *et al.* 1997). For this protein, the native secondary structure information was determined using the original PDB server, while the secondary structure constraints were predicted by the SCRATCH prediction server (Pollastri *et al.* 2002). By inspecting the Pareto front of 1ZDD protein (figure 5), we can note that there are no knees, hence we cannot use our decision-making method. In this case, we simply select

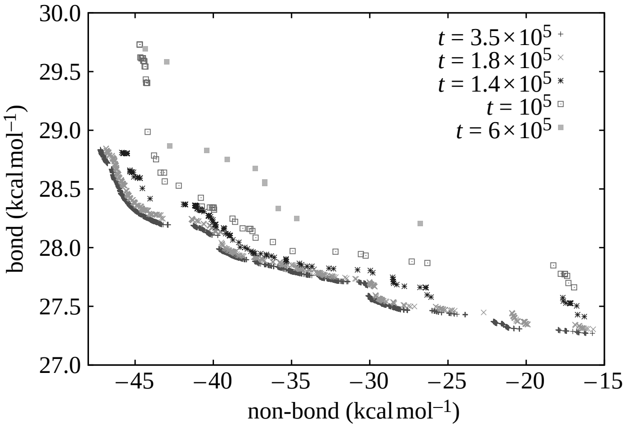

Figure 3. [Met]-enkephalin Pareto front dynamic (t is the number of iterations).

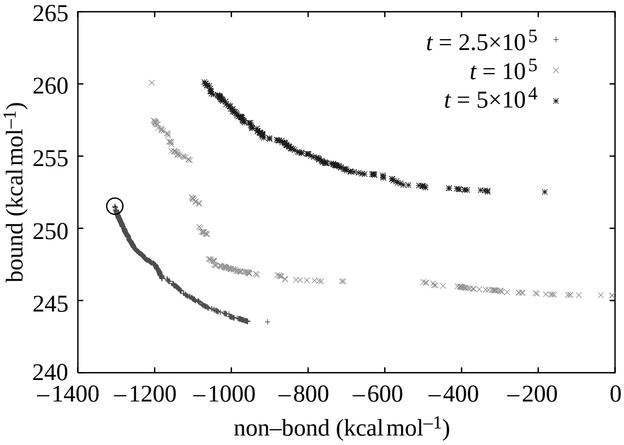

Figure 5. 1ZDD Pareto front dynamic (t is the number of iterations); the circle indicates the selected conformation.

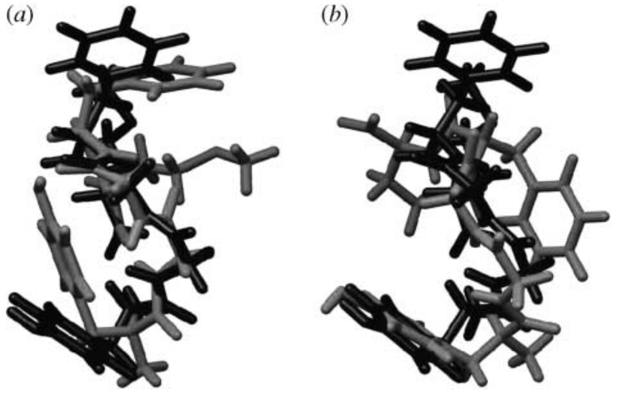

Figure 4. (*a*) Overlap between predicted (black) and Scheraga conformations for [Met]-enkephalin, $\text{DME}_{C_\alpha} = 0.454$ Å, $\text{RMSD}_{C_\alpha} = 0.490$ Å; (*b*) and overlap between predicted (black) and 1PLW conformations for [Met]-enkephalin, $\text{DME}_{C_\alpha} = 1.200$ Å, $\text{RMSD}_{C_\alpha} = 1.740$ Å.

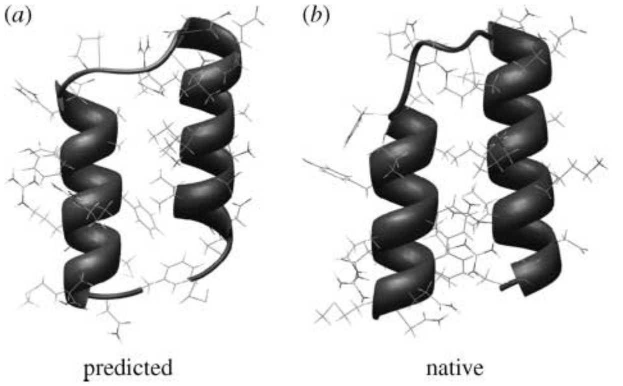

Figure 6. Predicted and native conformations for 1ZDD protein ($\text{DME}_{C_\alpha} = 1.54$ Å, $\text{RMSD}_{C_\alpha} = 2.27$ Å).

Table 3. Pareto front inspection results for 1ZDD protein. (The criteria used to select the best solution are given in bold.)

decision-making criteria	energy (kcal mol^{-1})	RMSD (Å)	DME (Å)
min energy	**−1052.09**	**2.27**	**1.54**
min RMSD	−1037.79	2.22	1.49

the conformation with lowest energy function value −1052.09 kcal mol^{-1}; this conformation matches the crystal structure with $\text{DME}_{C_\alpha} = 1.54$ Å and $\text{RMSD}_{C_\alpha} = 2.27$ Å (see figure 6). Table 3 shows the good relationship between the best energy structure and the best RMSD conformation in the final archive. Moreover, by inspecting the conformations in the final archive, we can see that they all present good characteristics both in terms of RMSD and energy. For this protein, the algorithm is able to produce an *ensemble* of good quality structures. Figure 7*a* shows the C_α RMSD per residue for the core region (3–32) of predicted structure. One of the two α-helix is better predicted than the other. In the plots, we also report the protein sequence, the predicted secondary structures for each residue and the native one.

Figure 7*b* displays a good correlation between the RMSD and the energy, suggesting that minimizing the energy by varying the conformation will tend to drive the conformation toward the true structure. Moreover, by inspecting the plot, it is evident that the algorithm is able to make a high sampling of the conformational search space: in a range of 1/2 Å there are more than 1000 conformations near the native state.

Repressor of primer (1ROP). Repressor of primer is a four-helix bundle protein that is composed of two identical monomers (Banner *et al.* 1987). Each monomer has 56 residues and forms α–turn–α structure (PDB id. 1ROP). For this protein the supersecondary structure constraints were predicted by the ANN method of Sun *et al.* (1997). The best computed structure, based on the decision-making method described in §5.2, matches the crystal structure with $\text{DME}_{C_\alpha} = 1.62$ Å, $\text{RMSD}_{C_\alpha} = 3.70$ Å and energy −797.57 kcal mol^{-1} (see figure 9). Table 4 shows the comparisons between the structures in the final archive based on different decision-making criteria.

In figure 8 we plot the observed Pareto front reporting many empty regions along the curve. These discontinuous regions show different clusters of non-dominated compact solutions near the folded state.

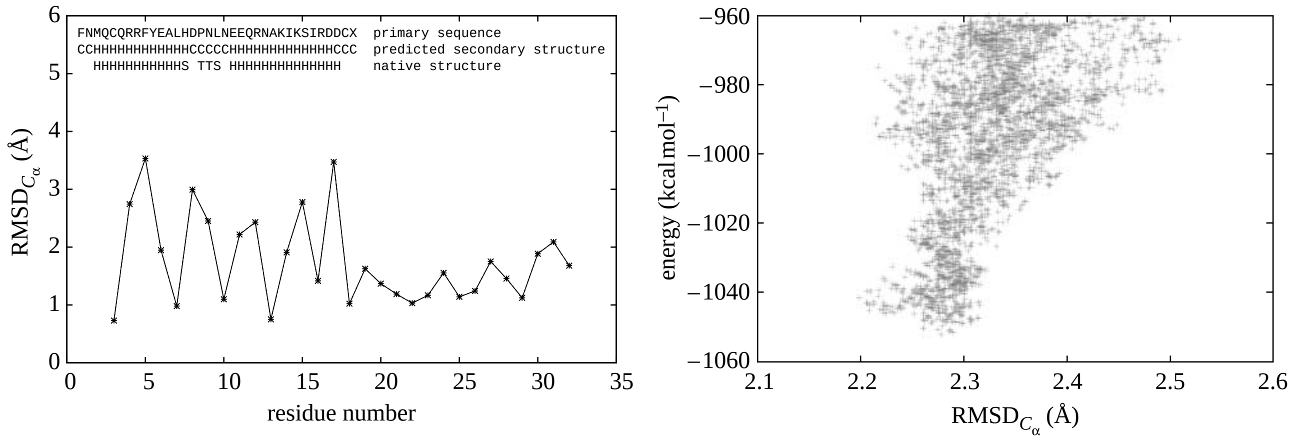

Figure 7. (*a*) C_α RMSD per residue for the core region (3–32) of 1ZDD protein. (*b*) Plot of the energy versus RMSD relative to 1ZDD protein for a set of more than 1000 conformations generated by the algorithm.

Table 4. Pareto front inspection results for 1ROP protein. (The criteria used to select the best solution are given in bold.)

decision-making criteria	energy (kcal mol^{-1})	RMSD (Å)	DME (Å)
min energy	−902.36	4.00	1.86
min energy in the knees	**−797.57**	**3.70**	**1.62**
min RMSD	−663.00	3.50	1.69

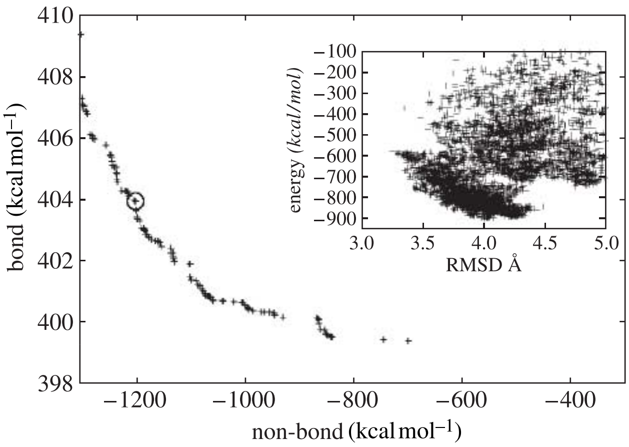

Figure 8. Pareto front and energy versus RMSD for 1ROP protein.

In the inset plot of figure 8 we show the correlation between the energy and the RMSD for conformations sampled by I-PAES algorithm. It is worth to note that it is possible to produce structures with lower energy than that of their native structures: a typical phenomenon observed in the search process of any method. In particular, this appears to be the case for the repressor of primer where the native structure energy (calculated with CHARMM) has value equal to −667.0515 kcal mol^{-1}, while the inset plot displays many native-like structures with lower energy values.

Uteroglobin (1UTG). Uteroglobin is a 4-helix protein that has 70 residues (Morize *et al.* 1987). The predicted supersecondary structures are α–bb–α–lbb–α–bb–α. Using these supersecondary structure constraints (predicted by the ANN method of Sun *et al.* 1997) the best computed structure, using our decision-making method described in §5.2, matches the native structure with DME$_{C_\alpha}$ = 3.79 Å, RMSD$_{C_\alpha}$ = 4.60 Å and energy 1128.3 kcal mol^{-1} (see table 5). The observed Pareto front of 1UTG obtained by the multi-objective evolutionary algorithm is a sparse set of points (see figure 11).

Crambin (1CRN). Crambin is a 46-residue protein with two α-helix and a pair of β-strands (Williams & Teeter 1984*a*,*b*). It has three disulphide bonds, whose constraints we do not use. The supersecondary structures, predicted by the ANN method of Sun *et al.* (1997), are β-loop–α–lbb–α–l–β-loop–α. The best computed structure, using our decision-making method described in §5.2, matches the crystal structure with DME$_{C_\alpha}$ = 3.72 Å, RMSD$_{C_\alpha}$ = 4.43 Å and energy 701.25 kcal mol^{-1} (see table 6). Figure 12 shows the relation between the Pareto front of the last iteration and the energy versus RMSD plot for 1CRN protein. Pareto front solutions are grouped into three individual clusters of non-dominated compact solutions. A wrong supersecondary structure prediction was made in the crambin at the C-terminal of the peptide chain: an incorrectly predicted α-helix (from residue 41 to 45) was imposed on the peptide chain as a constraint, this is evident from figure 13. Although, the wrong structure was formed in this terminal, the algorithm is able to reach a native like structure.

We would like to underline the fact that the protein conformation that has the minimum energy in the knees, is often better than the one from the whole obtained Pareto front (e.g. 1ROP, 1CRN and 1UTG proteins). Thus, as we mentioned above, the energy landscape produced by the CHARMM energy function does not seem to fit well the real landscape. Finally, the fact that solutions in the Pareto front that are not minimum energy solutions are better (in terms of RMSD and DME), clearly justifies a multi-objective approach (tables 4–6).

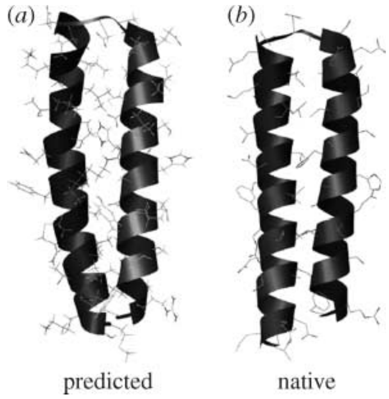

Figure 9. Predicted and native conformations for 1ROP protein ($\text{DME}_{C_\alpha} = 1.62$ Å, $\text{RMSD}_{C_\alpha} = 3.70$ Å).

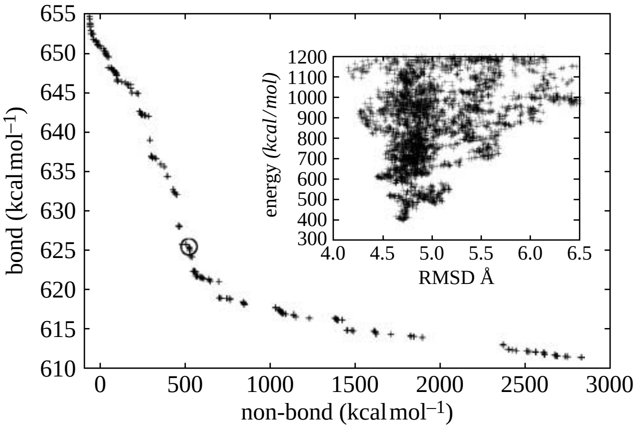

Figure 10. Pareto front and energy versus RMSD for 1UTG protein.

Table 5. Pareto front inspection results for 1UTG protein. (The criteria used to select the best solution are given in bold.)

decision-making criteria	energy (kcal mol^{-1})	RMSD (Å)	DME (Å)
min energy	573.89	5.31	4.52
min energy in the knees	**1128.3**	**4.60**	**3.79**
min RMSD	1170.85	4.27	3.47

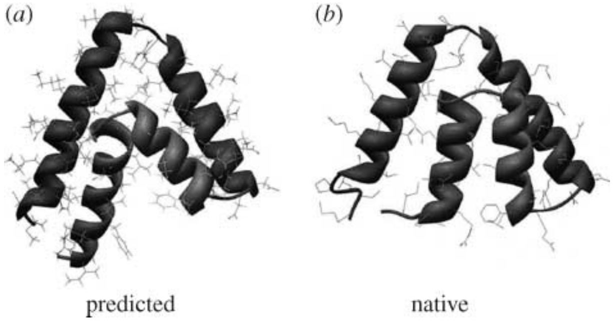

Figure 11. Predicted and native conformations for 1UTG protein ($\text{DME}_{C_\alpha} = 3.79$ Å, $\text{RMSD}_{C_\alpha} = 4.60$ Å).

Table 6. Pareto front inspection results for 1CRN protein. (The criteria used to select the best solution are given in bold.)

decision-making criteria	energy (kcal mol^{-1})	RMSD (Å)	DME (Å)
min energy	509.09	6.99	4.82
min energy in the knees	**701.25**	**4.43**	**3.72**
min RMSD	752.42	4.375	3.77

6.1. Comparisons with other approaches

We compared our algorithm, I-PAES, and its results to other works in literature (see table 8) and others MOEAs, in particular NSGA2 (Deb *et al.* 2002), that we implemented and tested on PSP. Two possible versions of NSGA2 were implemented. The first one uses standard low-level operators (SBX crossover and polynomial mutation; Deb 2001), and the protein is considered as a long sequence of torsion angles (real numbers). The second one uses high-level operators (naive crossover and the scheme of mutation used by I-PAES). In this case, the protein is manipulated at the amino acid level. The better performance of the high-level version is very clear, although the best RMSD found is always worse then that found by I-PAES. The best RMSD found for 1CRN by Cooper *et al.* (2003), using a Hill-climbing genetic algorithm, is 5.6 Å. Again, our method performed better in terms of best solution. Inspecting the results reported in table 9, I-PAES outperform also the good RMSD values obtained by the GA designed by Dandekar & Argos (1996). Table 9 shows the comparison between I-PAES, NSGA2, Hill-climbing GA (Cooper *et al.* 2003) and Dandekar & Argos' GA (1996) on 1CRN.

7. CONCLUSION

As reported by Plotkin & Onuchic (2002) 'the folded state is a small ensemble of conformational structures compared to the conformational entropy present in the unfolded ensemble'. This sentence characterizes our research goal of finding a set of equivalent three-dimensional conformations inside the folded state. To reach this goal we adopt a multi-objective approach in order to obtain good observed Pareto fronts of non-dominated compact solutions near or inside the folded state. We propose a modified version of the algorithm PAES that uses immune inspired principles (clonal expansion and hypermutation operators) as a new search method for PSP.[1]

The multi-objective approach is used to fold a peptide, the Met-enkephalin, and medium size proteins, and the results are comparable in terms of RMSD and DME to other approaches in the literature.

In the last 50 years, the PF and the PSP problems have been faced as a large single-objective optimization problem. In this article, we conjecture by computational experiments that, instead, it could be more

[1] The I-PAES source code is available from the authors.

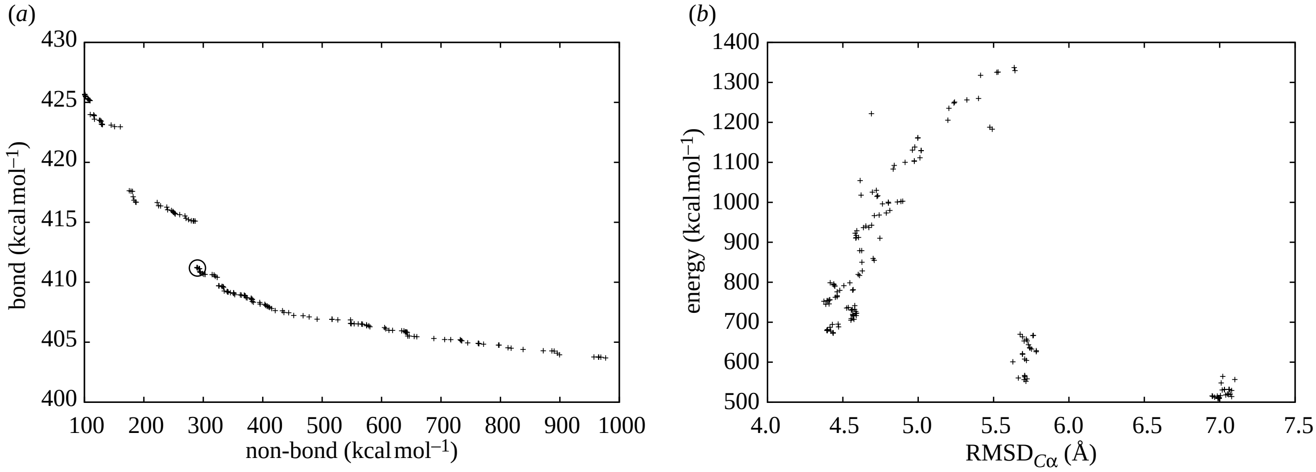

Figure 12. Comparison between the Pareto front (*a*) and energy versus RMSD values (*b*) for 1CRN protein.

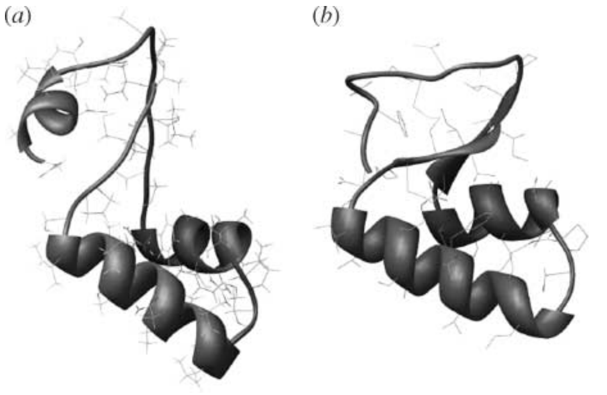

Figure 13. (*a*) Predicted and (*b*) native conformations for 1CRN protein ($\mathrm{DME}_{C_\alpha} = 3.72$ Å, $\mathrm{RMSD}_{C_\alpha} = 4.43$ Å).

Table 7. Protein results.

protein (PDB id)	ℓ	energy (kcal - mol^{-1})	DME_{C_α} (Å)	RMSD_{C_α} (Å)
Met-Enkephelin	5	−20.56	0.454	0.490
1ZDD	34	−1037.831	1.54	2.27
1ROP	56	−661.481	1.62	3.70
1UTG	70	1128.3	4.60	3.79
1CRN	46	410.038	3.72	4.43

suitable to model the PSP problem as a MOOP. Recently, the role of non-native and native interactions has been deeply studied (McLeish 2005). Specific yet non-native interactions may be important in stabilizing the low-dimensional diffusive searches on the folding pathways, as well as native interactions. The present research work considers the bond and non-bond interactions as main forces to direct the folding toward the native state. This model is based on the fact that local interaction (bond energy) and non-local interaction (non-bond energy) between atoms are in conflict.

Table 8. I-PAES versus other approaches for Met-enkephalin peptide.

algorithm	energy (kcal mol^{-1})	RMSD (Å)
I-PAES	−20.47±1.54 CHARMM	**2.835**
REGAL (real cod.) Tight constr. (Kaiser *et al.* 1997)	−23.55±1.69 CHARMM	3.23
Lamarkian (binary cod.) (Kaiser *et al.* 1997)	−28.35±1.29 CHARMM	3.33
Baldwinian (binary cod.) (Kaiser *et al.* 1997)	−22.57±1.62 CHARMM	3.96
REGAL (real cod.) Loose constr. (Kaiser *et al.* 1997)	−22.01±2.69 CHARMM	4.25
SGA (binary cod.) (Kaiser *et al.* 1997)	−22.58±1.57 CHARMM	4.51
REGAL (real cod.) (Kaiser *et al.* 1997)	−24.92±2.99 CHARMM	4.57

Table 9. I-PAES versus other approaches for 1CRN protein.

algorithm	RMSD (Å)
I-PAES	4.43
Dandekar & Argos' GA (1996)	5.4
HC-GA (with hydrophobic term; Cooper *et al.* 2003)	5.6
NSGA2 (with high-level operators)	6.447
HC–GA (no hydrophobic term; Cooper *et al.* 2003)	6.8
NSGA2 (with low-level operators)	10.34

It is clear that, although it is possible to make movements that locally are able to decrease the bond energy of the protein conformation, globally, this could be not true. Moreover, the electrostatic interactions or the short distance van der Waals interaction of atoms near in the space but far in the protein primary sequence could be penalized although an improvement is obtained in bond, angles or torsion energies. If one interaction decreases the other always increases but while this process goes on, there are compensation effects that will cause the minimization of the total energy. Experimentally, it has been shown that the two interaction types are in conflict following a typical characteristic of the MOOPs.

The authors would like to express their gratitude to the anonymous reviewers for their helpful comments. The computations were carried out on the Applied Computer Science Group Facility of the Ippari Research Center at the Comiso Campus of the University of Catania.

REFERENCES

Anfinsen, C. B. 1973 Principles that govern the folding of protein chains. *Science* **181**, 223–230.

Banner, D. W., Kokkinidis, M. & Tsernoglou, D. 1987 Structure of the ColE1 rop protein at 1.7 Å resolution. *J. Mol. Biol.* **196**, 657. (doi:10.1016/0022-2836(87)90039-8)

Bowie, J. U. & Eisemberg, D. 1994 An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl Acad. Sci. USA* **91**, 4436–4440.

Branke, J., Deb, K., Dierolf, H. & Osswald, M. 2004 Finding knees in multi-objective optimization. *LNCS* **3242**, 722–731.

Cooper, L. R., Corne, D. W. & Crabbe, M. J. 2003 Use of a novel Hill-climbing genetic algorithm in protein folding simulations. *Comput. Biol. Chem.* **27**, 575–580. (doi:10.1016/S1476-9271(03)00047-1)

Cornell, W. D. *et al.* 1995 A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197. (doi:10.1021/ja00124a002)

Cozzetto, D., Di Matteo, A. & Tramontano, A. 2005 Ten years of prediction … and counting. *FEBS J.* **272**, 881–882.

Cui, Y., Chen, R. S. & Wong, W. H. 1998 Protein folding simulation using genetic algorithm and supersecondary structure constraints. *Proteins: Struct. Funct. Genet.* **31**, 247–257. (doi:10.1002/(SICI)1097-0134(19980515)31:3<247::AID-PROT2>3.0.CO;2-G)

Cutello, V. & Nicosia, G. 2004 The clonal selection principle for *in silico* and *in vitro* computing. In *Recent developments in biologically inspired computing* (ed. L. N. de Castro & F. J. Von Zuben), pp. 104–145. Hershey, PA: Idea Group Publishing.

Cutello, V., Narzisi, G. & Nicosia, G. 2005 A class of Pareto archived evolution strategy algorithms using immune inspired operators for *ab initio* protein structure prediction. *Lect. Notes Comput. Sci.* **3449**, 54–63.

Dandekar, T. & Argos, P. 1996 Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J. Mol. Biol.* **256**, 645–660. (doi:10.1006/jmbi.1996.0115)

Day, R. O., Zydallis, J. B., Lamont, G. B. & Pachter, R. 2002 Solving the protein structure prediction problem through a multiobjective genetic algorithm. *Nanotechnology* **2**, 32–35.

Deb, K. 2001 *Multi-objective optimization using evolutionary algorithms.* Chichester, UK: Wiley.

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. 2002 A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comp.* **6**, 182–197. (doi:10.1109/4235.996017)

Dunbrack, R. L. & Cohen, F. E. 1997 Bayesian statistical analysis of protein sidechain rotamer preferences. *Protein Sci.* **6**, 1661–1681.

Foloppe, N. & MacKerell, A. D. 2000 All-atom empirical force field for nucleic acids. I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **21**, 86–104. (doi:10.1002/(SICI)1096-987X(20000130)21:2<86::AID-JCC2>3.0.CO;2-G)

Handl, J. & Knowles, J. 2005 Exploiting the trade-off—the benefits of multiple objectives in data clustering. Evolutionary Multi-Criterion Optimization (EMO 2005). *LNCS* **3410**, 547–560.

Hansmann, U. H. & Okamoto, Y. 1997 Numerical comparisons of three recently proposed algorithms in the protein folding problem. *J. Comput. Chem.* **18**, 920–933. (doi:10.1002/(SICI)1096-987X(199705)18:7<920::AID-JCC5>3.0.CO;2-T)

Hermans, J., Berendsen, H. J. C., van Gunsteren, W. F. & Postma, J. P. M. 1984 A consistent empirical potential for water–protein interactions. *Biopolymers* **23**, 1. (doi:10.1002/bip.360230807)

Jensen, M. T. 2003 *Guiding single-objective optimization using multi-objective methods.* Springer Lecture Notes in Computer Science, vol. 2611, pp. 268–279. Berlin: Springer.

Kaiser, Jr, C. E., Lamont, G. B., Merke, L. D., Gates, G. H. & Pachter, R. 1997 Polypeptide structure prediction: real-value versus binary hybrid genetic algorithms. In *Proceedings of the 1997 ACM Symposium on Applied Computing (ACM'97),* pp. 279–286. New York: ACM. (doi:10.1145/331697.331755)

Klepeis, J. L., Wei, Y., Hecht, M. H. & Floudas, C. A. 2005 *Ab initio* prediction of the three-dimensional structure of a *de novo* designed protein: a double-blind case study. *Proteins* **58**, 560–570. (doi:10.1002/prot.20338)

Knowles, J. D. & Corne, D. W. 1999 The Pareto archived evolution strategy: a new baseline algorithm for Pareto multiobjective optimisation. In *Proc. 1999 Congress on Evolutionary Computation* (ed. P. J. Angeline), vol. 1, pp. 98–105. Piscataway, NJ: IEEE Press.

Knowles, J. D., Watson, R. A. & Corne, D. W. 2001 Reducing local optima in single-objective problems by multi-objectivization. In *Proc. First International Conference on Evolutionary Multi-criterion Optimization (EMO'01)* (ed. E. Zitzler *et al.*), pp. 269–283. Berlin: Springer.

Li, Z. & Scheraga, H. A. 1988 Structure and free energy of complex thermodynamic systems. *J. Mol. Struct.* **179**, 333.

Louis, S. J. & Rawlins, G. J. E. 1993 Pareto optimality, ga-easiness and deception. *Proc. Fifth Int. Conf. Genet. Algorith.*, 118–123.

Ma, B., Kumar, S., Tsai, C.-J. & Nussinov, R. 1999 Folding funnels and binding mechanisms. *Protein Eng.* **12**, 713–720. (doi:10.1093/protein/12.9.713)

MacKerell Jr, A. D., Brooks, B., Brooks III, C. L., Nilsson, L., Roux, B., Won, Y. & Karplus, M. 1998 CHARMM: the energy function and its parameterization with an overview of the program. In *The encyclopedia of computational chemistry* (ed. P. v. R. Schleyer *et al.*), vol. 1, pp. 271–277. Chichester: Wiley.

McLachlan, A. D. 1982 Rapid comparison of protein structures. *Acta Cryst Allogr. A* **38**, 871–873. (doi:10.1107/S0567739482001806)

McLeish, T. C. B. 2005 Protein folding in high-dimensional spaces: hypergutters and the role of nonnative interactions. *Biophys. J.* **88**, 172–183. (doi:10.1529/biophysj.103.036616)

Mirny, L. A., Finkelstein, A. V. & Shakhnovich, E. I. 2000 Statistical significance of protein structure prediction by threading. *Proc. Natl Acad. Sci. USA* **97**, 9978–9983. (doi:10.1073/pnas.160271197)

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. 1975 Energy parameters in polypeptides VII, geometric parameters, partial charges, non-bonded interactions, hydrogen bond interactions and intrinsic torsional potentials for naturally occurring amino acids. *J. Phys. Chem.* **79**, 2361–2381. (doi:10.1021/j100589a006)

Morize, I., Surcouf, E., Vaney, M. C., Epelboin, Y., Buehner, M., Fridlansky, F., Milgrom, E. & Mornon, J. P. 1987 Refinement of the C222(1) crystal form of oxidized uteroglobin at 1.34 Å resolution. *J. Mol. Biol.* **194**, 725. (doi:10.1016/0022-2836(87)90250-6)

Nicosia, G. 2004 Immune algorithms for optimization and protein structure prediction. Ph.D. thesis, University of Catania, Italy.

Pendersen, J. T. & Moult, J. 1997 Protein folding simulations with genetic algorithms and a detailed molecular description. *J. Mol. Biol.* **169**, 240–259. (doi:10.1006/jmbi.1997.1010)

Plotkin, S. S. & Onuchic, J. N. 2002 Understanding protein folding with energy landscape theory. *Q. Rev. Biophys.* **35**, 111–167. (doi:10.1017/S0033583502003761)

Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. 2002 Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228–235. (doi:10.1002/prot.10082)

Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. 1997 Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring function. *J. Mol. Biol.* **306**, 1191–1199. (doi:10.1006/jmbi.2000.4459)

Starovasnik, M. A., Braisted, A. C. & Wells, J. A. 1997 Structural mimicry of a native protein by a minimized binding domain. *Proc. Natl Acad. Sci. USA* **94**, 10 080. (doi:10.1073/pnas.94.19.10080)

Steuer, R. E. 1986 *Multiple criteria optimization: theory, computation and application.* New York: Wiley.

Sun, Z. & Jang, B. J. 1996 Patterns and conformations commonly occurring supersecondary structures (basic motifs) in Protein Data Bank. *J. Protein Chem.* **15**, 675–690. (doi:10.1007/BF01886750)

Sun, Z., Rao, X., Peng, L. & Xu, D. 1997 Prediction of protein supersecondary structures based on the artificial neural network method. *Protein Eng.* **10**, 763–769. (doi:10.1093/protein/10.7.763)

Tramontano, A. & Morea, V. 2003 Assessment of homology based predictions in CASP 5. *Proteins* **52**, 352–368. (doi:10.1002/prot.10543)

Whisstock, J. C. & Lesk, A. M. 2003 Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* **36**, 307–340. (doi:10.1017/S0033583503003901)

Williams, R. W. & Teeter, M. M. 1984*a* Raman spectroscopy of homologous plant toxins: crambin and alpha 1- and beta-purothionin secondary structures, disulfide conformation, and tyrosine environment. *Biochemistry* **23**, 6796. (doi:10.1021/bi00321a080)

Williams, R. W. & Teeter, M. M. 1984*b* Raman spectroscopy of homologous plant toxins: crambin and alpha 1- and beta-purothionin secondary structures, disulfide conformation, and tyrosine environment. *Biochemistry* **23**, 6796. (doi:10.1021/bi00321a080)