Robust Bio-Active Peptide Prediction using Multi-Objective Optimization

Giuseppe Narzisi  
Courant Institute of Mathematical Sciences  
New York University  
New York, NY, 10012, USA  
Email: narzisi@nyu.edu

Giuseppe Nicosia and Giovanni Stracquadanio  
Department of Mathematics and Computer Science  
University of Catania  
Catania, 95125, Italy  
Email: {nicosia,stracquadanio}@dmi.unict.it

Abstract—Bio-active peptides control many important functions in organisms, such as cell reproduction, appetite, euphoria, sleep, learning, immune response, etc. They also act on hormones, neurotransmitters, antioxidants, toxins and antibiotics. Because of their importance, bioactive peptides have received particular attention and extensive studies have been carried out to determine their structures. Although their typical size does not exceed 30 amino acids, their 3D structure is challenging to predict because of the following reasons: (i) their conformation often includes β-turns which are more difficult to predict using standard potential energy functions; (ii) they fold into structures that are not similar to already known proteins, which makes them hard instances for comparative modeling techniques; (iii) they are more exposed to the solvent than longer proteins and this additional effect has a consequence on their final conformation.

This paper presents a strategy for peptides structure prediction that uses: (1) a multi-objective formulation of the optimization problem, (2) a multi-objective evolutionary algorithm to explore the search space, (3) a decision making phase based on different metrics to select solution from the Pareto front, and (4) a method to analyze the robustness of the solution using the Monte Carlo method. We have tested this prediction pipeline on a large dataset of 43 bioactive peptides and the experimental results show that this method outperforms the PEPstr prediction server and is competitive against a more recent Generalized Pattern Search approach. Multiple solutions can be generated, as opposed to standard single-objective methods, which are generally more robust than the wild-type.

Keywords—Peptides Prediction; Multi-Objective Optimization; Monte Carlo Robustness; Decision Making Strategies;

I. INTRODUCTION

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, however it is only after proteins assume their native 3D structure that they can carry out their biochemical function. The pioneering experiments of Anfinsen in 1973 [1] shed light on the mechanism involved in protein folding. According to the thermodynamic hypothesis, proteins are not assembled into their native structures by biological process, but folding is a physical process that depends only on the specific amino acid sequence and the surrounding solvent. This implies that the native state of a protein is the state of lowest free energy of the protein system under physiological conditions. Based on this observation, the Protein Structure Prediction problem (PSP) has been cast into an optimization problems that consists of identifying the lowest free energy structure of the protein in its environment using only laws of physics and the amino acid sequence. This approach is referred to as ab initio because no background knowledge on experimentally known structures is used during the prediction.

Historically the protein structure prediction has been approached as single-objective optimization problem: given the primary sequence of the protein, find the 3D native conformation with minimum energy using a single-objective potential energy function. However, recent advances in protein structure prediction have shown that, due to multiple conflicting interactions between atoms and the low accuracy of current potential energy functions, it is convenient to tackle the problem as a multi-objective one [2]. Common 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). These two types of interactions have been shown experimentally to be in conflict. A multi-objective evolutionary algorithm (I-PAES) is used in [2] as a search procedure for exploring the conformational space of the PSP problem. This approach has been effective for medium-sized proteins (up to 70 residues) and a recent review paper published by the Journal of the Royal Society Interface [3], has ranked I-PAES among the best state-of-the-art folding algorithms. In this paper we further explore the capabilities of this multi-objective framework by testing it on a large test-bed of shorter protein sequences (up to 20 amino acids) that control many functions of living organism: bio-active peptides. Despite their short size, these proteins are particularly challenging because they have no regular secondary structure and they do not fold into structures similar to that of already known proteins.

The paper is organized as follows. The method section introduces all the required tools and algorithms: section II-A describes the benchmark, sections II-B and II-C define the energy function and its multi-objective decomposition, section II-D lists the quality metrics, section II-E describes the computation of the solvent energy, section II-F shows...
the users served for secondary structure prediction, section II-G briefly discusses the I-PAES algorithm and section II-I puts together all the tools into the prediction pipeline. The results section shows the performance of the multi-objective framework on the 43 peptides benchmark and the comparison to the state-of-the-art (section III-A). Multiple decision-making strategies are proposed (in section II-H) and evaluated based on different quality metrics and robustness of the solutions in section III-B. Finally, the conclusion section discusses the results and brings the paper to a close.

II. MATERIALS AND METHODS

A. The PEPstr benchmark

The protein benchmark consists of a large test bed of 43 proteins proposed by Kaur et al. [4]. A collection of 77 bioactive proteins were selected from the Protein Data Bank and other resources such as PRF1. The dataset was restricted only to peptides that consist of natural amino acids with length between 9 and 20 residues. Peptides stabilized by a disulfide bridge were excluded (as proposed in [4]). After these refinement processes, the peptides were divided into three groups according to the secondary structure elements: 32.2% helical, 6.9% β-sheet, 34.9% β-turns. Table I shows the PDB code, class, length, and energy information for each peptide in the benchmark. Kaur et al. in [4] propose a new method, called PEPstr, to predict the 3D structure of these proteins. In particular they build four different structural models for each peptide and test their performance. In the first model an extended conformations is used for each residue (Φ = ψ = 180°); the second model uses secondary structure information predicted from PSIPRED [5]; in the third model the secondary structure information for β-turns types is also used and predicted using PROMOTIF [6], in the fourth model the side chain angles are introduced and set according to the standard Dunbrack backbone dependent rotamer library [7].

B. Empirical Conformational Energy Program for Peptides

The ECEPP potential energy function is a well known force field for proteins developed by F.A. Momany, H.A. Scheraga and colleagues [8]. ECEPP was developed specifically for modeling of peptides and proteins. It uses fixed geometries of amino acid residues to simplify the potential energy surface. Thus, the energy minimization is conducted in the search space of the torsion angles. In the ECEPP force field the internal potential consists of two parts: the spatial energy and the torsion energy. The spatial energy is expressed as the sum of the pairwise interactions between atoms, which includes the electrostatic term, Lennard-Jones attractive and repulsive term, and the hydrogen bond potential. The torsion energy is simply reduced to a sum of the torsion energy for each bond, which is solely a function of the corresponding dihedral angles. The initial ECEPP formulation has been modified and improved over the years, in this paper the following ECEPP/3 formulation is used:

\[
E_{\text{ecepp}} = \sum_{(i,j)} 332q_i q_j \varepsilon r_{ij} + \sum_{(i,j)} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \sum_{(i,j)} \left( C_{ij} r_{ij} - D_{ij} \right) + \sum_{n} U_n [1 \pm \cos(k_n \phi_n)]
\]

where, \(r_{ij}\) is the distance in Å between atoms \(i\) and \(j\); \(A_{ij}, B_{ij}, C_{ij}\) and \(D_{ij}\) are parameters of the empirical potentials; \(q_i\) and \(q_j\) are the partial charges on the atoms \(i\) and \(j\), respectively; \(\varepsilon\) is the dielectric constant of environment;

\[1\text{http://www.genome.ad.jp}\]
$U_n$ is the energetic torsion barrier of rotation about the bond $n$; $k_n$ is the multiplicity of the torsion angle $\phi_n$.

C. ECEPP multi-objective formulation

The ECEPP force field can be decomposed in two partial sums reflecting bonded and non-bonded atom energies. Thus, based on the ECEPP/3 equation (1), the two objectives to minimize become:

$$f_1 = E_{\text{bond}}(A_{\text{bond}}, C_{\text{bond}}) = E_{\text{tor}}$$
$$f_2 = E_{-\text{bond}}(A_{-\text{bond}}, C_{-\text{bond}}) = E_C + E_{\text{LJ}} + E_{\text{HB}}$$

where symbols $C_{\text{bond}}$ and $C_{-\text{bond}}$ represent respectively the force constants involved for bond and non-bond atoms in equation (1), Symbol $A_{\text{bond}}$ represents the set of all atom chains of max length four (bonds, angles and torsion interactions between atoms), and symbol $A_{-\text{bond}}$ represents all the atoms not connected by chemical bond. For all the results presented in this paper we used the SMMP$^2$ [9] (Simple Molecular Mechanics for Proteins) program to compute the energy of a protein in the ECEPP model.

D. Quality Metrics

Different quality metrics have been proposed to compare the 3D conformation of two proteins. Here we describe the two that we use in the results.

Root Mean Square Deviation: The root mean square deviation (RMSD) is the measure of the average distance between the backbones of superimposed proteins after optimal rigid body superposition and it is defined as follows:

$$\text{RMSD}(A, B) = \sqrt{\frac{\sum_{i=1}^{L_N} (|a_i - T(b_j)|)^2}{L_N}}$$

where $T(\cdot)$ is the rigid transformation that optimally superimposes the two structures $A$ and $B$, $a_i$ and $b_i$ are the 3D coordinates of atoms $i$ in $A$ and $B$ respectively, and $L_N$ is the length of the protein. The fitting is usually performed using the McLachlan algorithm [10], which is based on a fast conjugate gradients method of superposing two sets of atomic coordinates.

Distance Matrix Error: The distance matrix error (DME) is a measure of the similarity between inter-atomic distances, and it is defined as follows:

$$\text{DME}(A, B) = \frac{2}{L_N(L_N - 1)} \sqrt{\sum_{i,j} (d_{(ij)} - d'_{(ij)})^2}$$

where $d_{(ij)} = ||a_i - a_j||$ and $d'_{(ij)} = ||b_i - b_j||$ are the distances between atoms $i$ and $j$ in structure $A$ and $B$ respectively. The calculation of the DME between two structure does not require the superposition of the coordinates. RMSD, which measures the similarity of atomic position, is usually larger than the DME, which measures the similarity between inter-atomic distances.

E. Solvent

In nature proteins exist in an environment of water and solvated salts. Due to their short sequence bio-active peptides are typically more exposed to the solvent. Electrostatic interactions play an important role on the behavior of the protein-solvent system [11]. The exact calculation of such interactions is computationally very expensive because of the high number of degrees of freedom of the protein-water system. The typical approach to reduce such complexity is to compute the solvent-accessible area, where it is assumed that the free energy contributions from atomic groups immersed in the protein interior differ from contributions of groups exposed to the water [12]. It is commonly accepted that this free energy difference is proportional to the surface area of the atomic groups exposed to the solvent. Using this approximation the total solvent energy of the protein is given by the formula $E_{\text{sol}} = \sum_i \sigma_i A_i$, where the sum is extended over all solvated atomic groups, $A_i$ is the conformational dependent solvent accessible area of the surface of atom $i$ and $\sigma_i$ is the atomic solvation parameter for atom $i$. Although the method is widely used, it has some drawbacks. Except for the question if the solvent-accessible area represents a good estimate of the solvent-protein interaction, the method suffers from the additional problem that the choice of the parameters is not simple and many authors have evaluated these parameters with different methods. For example the SMMP package offers 9 different sets of solvation parameters. We use the SMMP program to compute the solvation energy for a protein conformation using the parameter set 1 from the package.

F. Secondary structure prediction

For protein sequence of length less than 15 residues, we use no secondary structure information, so all the torsion angles are free to move in the whole range $[-180^\circ, 180^\circ]$. Only for peptides with more than 15 residues the secondary structure information is predicted and used to constrain the angle values. In particular the bounds for the main chain ($\phi, \psi$) are set according to Klepeis et al [13]. Two different servers have been used for prediction: SCRATCH [14] and Porter [15]. We observed poor performance of the SCRATCH server for the prediction of small peptides. In fact, for some of the instances of the PEPstr benchmark the SCRATCH produces wrong predictions for relatively long subsequence in the peptides. These kind of errors are more significative in a short peptide than they would be in a longer protein, since they could completely change the final shape of the predicted 3D structure. For some of the peptides where the SCRATCH server fails we have used the Porter server which has shown better performance. The predicted secondary structures using SCRATCH and Porter servers are listed in the supplementary material.$^3$

$^2$http://www.smmp05.net/

$^3$http://cims.nyu.edu/~gn387/peptides_prediction.pdf
G. I-PAES algorithm

In this work we employ the multi-objective evolutionary algorithm proposed in [2] as the search procedure for exploring the conformational space of the protein. I-PAES is a modified version of PAES [16] with a different solution representation (polypeptide chain) and immune inspired (cloning and hypermutation) operators (see paper [2] for the pseudo-code and description of the algorithm).

H. Decision Making (DM) strategies

Since the solution of a multi-objective optimization problem is composed by a set of non-dominated solutions, it can be difficult to select one single solution from the Pareto set. There exist different ways to select a solution from the Pareto front and many of them are based on simple geometric arguments, however there is no comparative study on their performance for general multi-objective optimization and in particular for protein structure prediction. As we tested the multi-objective approach on such a large set of proteins we have also experimentally checked the performance of these selection techniques. The description of the different DMs used in the results is reported next while Figure 1 gives their geometric interpretation.

Closest to Ideal: Since we are minimizing in parallel different objectives, the most obvious approach is to select the solution that is the closest to the point corresponding to the minimum of all the objectives. Such a solution is called ideal objective vector and it is defined as following:

\[ z^* = (f_1^*, f_2^*, \ldots, f_M^*)^T \]  

(6)

where \( f_i^* \) is the function value associated with the minimum solution \((x_i^*)\) for the \( m \)-th objective function. Figure 1 shows both the ideal and close to ideal solutions.

\[ \text{Min} \]

Lower bounds: Based on definition (6), we can select each solution representing the lower bound for each of the objectives. In the specific case of the 2-objective formulation given here, this strategy corresponds to selecting the two solutions having minimal bond and non-bond energies. These solutions are labeled respectively \( \text{Min}_{\text{Bond}} \) and \( \text{Min}_{\text{Non-Bond}} \) in Figure 1.

Angle based: Typically the most interesting solutions of the observed Pareto front are solutions lying on the knees of the front [2], [17]. For these solutions small improvement in one objective will cause a large deterioration in at least one other objective. A simple way to identify a knee-solutions is based on the selection of a fixed number of closest neighbors around each solution and compute the angles spanning between them. The max of these angles is assigned to the central solution. In this work we fix the number of neighbors to 4 (2 solution per side) whose combination produces 4 different angles to evaluate. The angle based solution is labeled \( \text{Max}_{\text{angles}} \) in Figure 1

External quality measures: Finally, additional quality measures can be used to select solution from the Pareto front (for instance structural robustness, compactness, hydrophobic score, etc.). In this study we employ the solvent energy, based on the computation of the solvent-accessible area (described in section II-E), as an additional quality measure to chose solution from the Pareto front.

I. Pipeline

The structure prediction of a protein is accomplished in several steps: 1) secondary structure prediction (SCRATCH or Porter); 2) Pareto front computation using I-PAES; 3) selection of solutions from the Pareto front using the DM strategies described in Section II-H; 4) Monte Carlo robustness analysis of the selected structures (see Section III-B). Finally the users can select a single solution based on his/her preference of robustness, energy, compactness, etc.

III. RESULTS

A. Results on the PEPstr dataset

This section shows and discusses the results obtained by I-PAES on the PEPstr benchmark (see Table II). The Table reports the best results in terms of \( DME \) and \( RMSD \) out of 10 independent runs. Also, \( RMSD \), energy and solvation are listed for the conformation with minimal energy in the final Pareto front at the end of \( 2.5 \times 10^5 \) number of energy function evaluations.

Almost all the single-objective optimization approaches for the PSP suffer from over-fitting the potential energy of the predicted structures, producing structures with energy less than the native ones, and, in fact, it is not clear if such behavior is useful during the search process. However, the multi-objective approach applied here is able to generate good quality non-dominated conformations with potential energies not significantly below the value of the native
protein. In fact, if we compare the energy values of the predicted conformation with the native ones (in Table I) it can be observed that they are very similar in the majority of the instances. The predicted conformations are also properly exposed to the solvent as reflected by the solvation energies term which is very close to the native one for each peptide. This result appears very unexpected to us, since the solvent-accessible area was not directly used during the optimization process but only evaluated at the end of the simulation.

Figure 2 compares the average results on 10 independent runs of the RMSD of solutions in the Pareto front selected according to different decision-making strategies: (1) closest to ideal, (2) min solvation energy, (3) min non-bond energy, (4) min bond energy, (5) max 4-angles value. Inspecting the plot, it seems that no particular selection strategies outperform any other, and that for the peptide benchmark there is no dominant strategy. This can be due to the small length and low complexity classes of the proteins in the test-bed. In fact, for these proteins, either the algorithm generates a very good Pareto front of solutions in terms of all the metrics or it wrongly predicts their structures.

If we compare the results obtained by I-PAES with the PEPstr approach (Table III) we notice that the multi-objective method outperforms the PEPstr model IV even though no secondary information is used for peptides of length less the 15 residues. Also the approach is very robust since the average RMSD on 10 independent runs are of the same order of the best results for PEPstr. We also compare I-PEAS with GPS [18]. GPS is a Generalized Pattern Search Algorithm based on the well known class of Search-and-Poll algorithms and it has been recently applied successfully to the structure prediction problem of the PEPstr peptide benchmark. GPS has the best performance on the proposed testbed but, since it is a single-objective approach, it returns as output only the best energy conformation for the protein.

Finally, if min RMSD solutions are selected from the Pareto front, the average result, as reported in Table III, is 2.21 Å. Even if this selection criterion is not legal, since it involves the comparison with the native structure, which is not available for newly discovered proteins, it is indicative of the quality of the solutions in the final Pareto front.

### B. Robustness Analysis

Robustness is believed to be one of the main driving forces of the evolution of living organisms; the ability of a biological system to survive both endogenous and exogenous mutations is crucial in introducing biological diversity. The idea of identifying robust systems is well established in system engineering, however this theoretical foundation can be also applied to control the robustness of living systems, as recently proposed by Kitano in [19], and the same principle can be applied in the context of protein structure prediction [20]. In particular, robustness is related to the ability of a protein to remain in its native state, thus preserving its function, even if it is subject to atomic perturbations. We define a protein to be robust if its potential energy has a tight variation under molecular perturbations.

Following the framework proposed in [20], in order to assess the robustness of our predictions, we adopt the...
energetic yield metric defined as follows:

$$EY(A^p_{MC}, P_{ref}, \epsilon) = \left| \frac{|P|_i \in A^p_{MC} : |E(P_i) - E(P_{ref})| < \epsilon}{|A^p_{MC}|} \right|$$

(7)

where $A^p_{MC}$ is the ensemble of conformations generated by the $p$ perturbation; since the energetic yield depends on the applied perturbation, it can be denoted also as $p$-yield or $p$-analysis. This metric evaluates the percentage of perturbed conformations that differs at most $\epsilon$ kcal/mol from the energy of the reference structure; according to [20], we fixed $\epsilon = 1.0$ kcal/mol. In order to generate an ensemble of perturbed conformations, we adopt a Monte-Carlo sampling algorithm; this method keeps in input a reference conformation $P_{ref}$, and generates a set of conformations applying a perturbation $p$. In our research work, we assume that perturbations can be represented by a standard normal distribution; we apply this noise on the torsion angle of the $P_{ref}$ structure, although we assure that each angle belongs to the interval $[-180^\circ, 180^\circ]$. We take into account three types of perturbations; Global, Local and Residue mutations. Global perturbation perturbs all the dihedral angles of the structures; this kind of mutation is designed to study dramatic changes in the polypeptide chain. Local perturbation applies a noise to one torsion angle at time; this one-factor-at-time analysis aims to discover sub-optimal atomic location in the prediction. The residue perturbation mutates each torsion angle of one amino-acid at a time; this robustness analysis is able to discover sensitive residues in the polypeptide chain, that are responsible for a low quality prediction or structural instability.

As an example, we investigate the robustness of the 1C98 chosen by the six decision making strategies; for each structure, we generate $10^4$ perturbed conformations for the global analysis, and 200 decoys for each torsion angle and aminoacid for the local and residue analysis, respectively.

The experimental results are presented in Table IV, where two main results are evident: (i) I-PAES is able to predict structures that are typically more robust than the wild-type, and this property seems to be relatively independent from the decision strategy; (ii) global yield is a good metric to select the most robust peptides and in particular best-bond and min-solvent strategies produce the most promising predictions. This result matches with the nature of the instances; peptides are typically fully exposed chains, and hence it is plausible that structures that maximize solvent contacts are robust.

The histograms in Figure 3 show the distributions of solutions, according to potential energy, of the most robust (best-bond) and the most sensitive (max-4-angles) predictions. The high yield of the best-bond solution implies a well centered bell around the energy value of the reference structure, conversely the max-4-angles conformation shows a more dispersed distribution not centered around the reference value. Also note that the energy range for the best-bound solution is tighter than the max-4-angles.

### IV. Conclusion

This paper has presented an extensive analysis of a new multi-objective prediction pipeline for the PSP (based on the work presented in [2]) on a benchmark of 43 bioactive peptides using a multi-objective formulation of the ECEPP energy function. The proposed method outperforms the PEPstr prediction server [4], although no secondary structure prediction is used for proteins of length less than 15 residues, and it is comparable to the GPS approach recently presented in [18]. Several decision-making strategies have
been formulated and exploited for the selection of solutions from the Pareto front generated by I-PAES. In the case of the 43-peptides dataset, choosing the solution with minimal energy has been the best strategy for characterizing conformations with low RMSD values and solvation energy close to the native one. However structural similarity to the native protein is not the only metric that characterizes good quality solutions. A solution close to the native state should also be robust: small variation to the torsion angles of the structure should not produce large variations in energy value. Monte Carlo Analysis has been employed to discern between robust and non-robust conformations from the Pareto front. Experimental results on the 1C98 peptide have shown that I-PAES generates conformations that are more robust of the wild-type independently from the decision strategy used. Both min-solvent and min-bond DMs are good selection strategy when robustness and stability are required.

Future work will focus to develop novel decision-making strategies that are able to identify native-like solutions in the Pareto front. Together with the geometric arguments presented in this paper, additional quality measure must be considered. For instance, structure stability, compactness and hydrophobic score are all major forces that play an important role in the folding process of a protein and their intelligent combination could generate new decision-making tools. Nevertheless, there are other missing elements from chemistry and biology that needs to be first discovered to improve the accuracy of current potential energy functions. The multi-objective prediction pipeline, with the decision-making phase at the end of the optimization process, allows to partially overcome this current lack of knowledge.

REFERENCES


