Full-atom ab initio protein structure prediction with a Genetic Algorithm using a similarity-based surrogate model

Conference Paper · July 2010

DOI: 10.1109/CEC.2010.5585959 · Source: DBLP

CITATIONS
13

READS
627

3 authors:

Fábio Lima Custódio
Laboratório Nacional de Computação Científica
33 PUBLICATIONS **358** CITATIONS
SEE PROFILE

Helio Barbosa
Laboratório Nacional de Computação Científica
219 PUBLICATIONS **3,845** CITATIONS
SEE PROFILE

Laurent E Dardenne
Laboratório Nacional de Computação Científica
84 PUBLICATIONS **2,529** CITATIONS
SEE PROFILE

Full-Atom *Ab Initio* Protein Structure Prediction with a Genetic Algorithm using a Similarity-based Surrogate Model

Fábio L. Custódio, Hélio J. C. Barbosa, *Member, IEEE*, and Laurent E. Dardenne

***Abstract*— The protein structure prediction problem is one of the most interesting challenges of computational biology. One of its critical facets is the optimization method employed. This is often carried out by metaheuristics, such as Genetic Algorithms (GA). The prediction involves optimization of a complex and computationally expensive energy function. Thus, the usual GA requirements of a large number of function evaluations can ultimately result in prohibitive computational costs. We applied a *k*-nearest neighbors surrogate modeling strategy, with two different similarity criteria, to improve the quality of proteins structures predicted by a crowding-based steady-state GA, without increasing the number of exact fitness evaluations. Additional protein conformations can be investigated using the surrogate model, potentially increasing the exploratory capability of the algorithm. The results obtained from six test proteins suggest that the surrogate model approach has the potential to improve the performance of the described protein structure prediction method.**

I. Introduction

It is a common practice to employ metaheuristics to investigate many interesting problems formulated as optimization problems in science and engineering. The complexity of these problems has been steadily increasing and thus resulting in more realistic models that, in turn, translate into computationally expensive simulations. One particular area where this phenomenon is notorious is molecular modeling [1].

The protein structure prediction (PSP) problem, is one of the most interesting challenges of modern computational biology [2]. The problem consists in determining the native conformation of proteins, i.e., given an amino acid sequence determine the 3D structure – the *native structure*. Methods for PSP have important biotechnological applications, e.g., the creation of new proteins (*de novo* protein design) [3], aiding structure based drug design projects (receptor flexibility) [4], refinement of theoretical models obtained by comparative modelling [5], and obtaining experimental structures from incomplete Nuclear Magnetic Resonance data [6].

Most PSP methods follow the thermodynamics hypothesis, i.e., conformations associated with global minima of an energy function are considered the native structure. The conformation the protein adopts under physiological conditions is the conformation with the lowest Gibbs free energy [7]. The PSP problem can be decomposed in two sub-problems: (i) to define an appropriate energy function that places the native structure on the global minimum and is able to discriminate correct from incorrect folds, and (ii) to develop an efficient and robust search strategy. The prediction involves optimization of a computationally expensive energy function with thousands of degrees of freedom associated with extremely complex energy hypersurfaces, that is, highly degenerated (including multiple minima), with massive multi-modality (roughness), and large regions of unfeasible conformations. This makes the problem difficult to treat, nevertheless rewarding, since the development of techniques capable of effectively dealing with such difficulties brings innovations applicable to a wide range of problems, especially closely related problems, such as the study of folding pathways and molecular docking.

The search for a method capable of predicting the correct structure, in the absence of known reference structures, is still an open problem. Recent efforts are periodically analyzed with protein sequences of elucidated, but unpublished, structures during the CASP (Critical Assessment of Structure Prediction) meetings. Currently the most promising methods utilize information of known structures, although they are still based on optimizing complex and expensive energy functions [8]–[10]. This optimization is often carried out by metaheuristics and, amongst them, Genetic Algorithms (GA) are noteworthy [11]. GA's robustness and wide applicability can be explained, at least partially, by their stochastic nature, and because they work with a population of candidate solutions (that makes the method naturally parallel). Other attractive feature is that they do not require differentiability or continuity of the fitness function. Despite their advantages, they usually require a large number of fitness function evaluations in order to reach optimal or near optimal solutions, and when expensive models are involved, such as the models used for PSP with atomic details, strategies for obtaining satisfactory solutions using viable computational resources are needed.

One possible solution to this problem is the use of approximations of the fitness function, that is, to use a computationally inexpensive surrogate model in place of the fitness function. Surrogate models are based on approximation/interpolation methods, statistical models, and, in some cases, *ad hoc* constructions from the original model. The most commonly used surrogate model methods are the Response Surface Methodology [12], Radial Basis Functions [13], [14], Artificial Neural Networks [15], Kriging or Gaussian Processes [16] and Support Vector Machines [17].

Fábio L. Custódio is a post-Doc at the Department of Applied and Computational Mathematics, Laboratório Nacional de Computação Científica, Petrópolis, Brazil; email: flc@lncc.br.

Helio J. C. Barbosa is with the Department of Applied and Computational Mathematics, Laboratório Nacional de Computação Científica, Petrópolis, Brazil; email: hcbm@lncc.br.

Laurent E. Dardenne is with the Department of Computational Mechanics, Laboratório Nacional de Computação Científica, Petrópolis, Brazil; email: dardenne@lncc.br.

978-1-4244-8126-2/10/$26.00

The main advantage in using a surrogate model is the decrease in computational time required in the evaluation of each individual in the population, although some studies report the facilitation of the optimization procedure by smoothing the objective function landscape [18], [19]. For a GA with a fixed computational budget, that could translate into additional function evaluations being performed and thus a more thorough exploration of the fitness landscape, potentially leading to better solutions.

In this paper we apply a similarity-based surrogate model, the Nearest Neighbor approximation model [20], to assist a crowding-based steady-state GA (CSSGA) for the PSP problem [21]. The model used here has been previously applied to a range of test functions and resulted in improvements on the performance of the genetic algorithm [22].

The paper is organized as follows. Section II gives a description of the PSP problem, the CSSGA and the details of the implementation of the surrogate model. Section III presents the results and is followed by the discussion.

II. Genetic Algorithm for Protein Structure Prediction

A. The Protein Structure Prediction Problem

The search for an appropriate fitness function for problems that involve modeling of macromolecules, including the PSP and the protein-ligand docking, has led to the development of functions that make use of several approaches. Most of these functions use simple models of potential energy functions based on force fields from classical molecular mechanics, without the explicit consideration of entropic effects [23]. Thus, the total energy of a protein structure is given by the energy of interaction between their atoms, and this is calculated as the combination of the energy resulting from: the electrostatic interactions, the atomic repulsion at very small distances, the attraction of van der Waals (vdW) and the energy barriers from the rotation of bonds. These energies are, respectively, modeled by the Coulomb potential, the Lennard-Jones potential, and the proper dihedral potential. The best conformational energy is the minimum of the energy function (E), where, despite the errors introduced by using a simplified energy model, is believed to be the native structure of the protein. Ideally the PSP problem can be formulated as:

$$\text{minimize } E(r_1, r_2, \ldots, r_N)$$

where N is the number of atoms and r_i are vectors containing the atomic Cartesian coordinates.

B. Representation of solutions

Although it seems natural to encode protein structures in chromosomes containing the Cartesian coordinates of each atom, this would not be practical for the PSP problem. Changes in the structure, by the application of the genetic operators, would often generate conformations with steric collisions with unnatural bond geometries. To deal with these distortions it would be necessary to use a repair mechanism. This, in turn, would be computationally very expensive [24].

In a real protein the atoms cannot move as free particles because they are connected, and these connections follow specific geometries (according to the types of atoms participating in the link). Since the atomic composition of the protein does not change during the PSP, these geometries can be automatically fulfilled by using an internal coordinates representation [25] and performing movements in the structure by changing the torsion angles. In the real system, the distances and angles of covalent bonds are not fixed, but the changes that occur around the equilibrium values are usually small. Thus, an approach commonly used by PSP methods is to keep the geometry of chemical bonds fixed during the search.

The representation used here is based on internal coordinates. The conformation of a peptide chain can be defined by a series of torsion angles, called backbone dihedral angles ϕ, ψ and ω, as shown in Figure 1(a). That means we can store the protein structures on a chromosome containing the values of those angles for each amino acid in the sequence. The positioning of side chains is also defined by a set of torsion angles: χ_1, χ_2, χ_3 and χ_4, as exemplified in Figure 1(b). However, these are not directly modified by the GA; instead they are defined according to a backbone-dependent rotamer library [26].

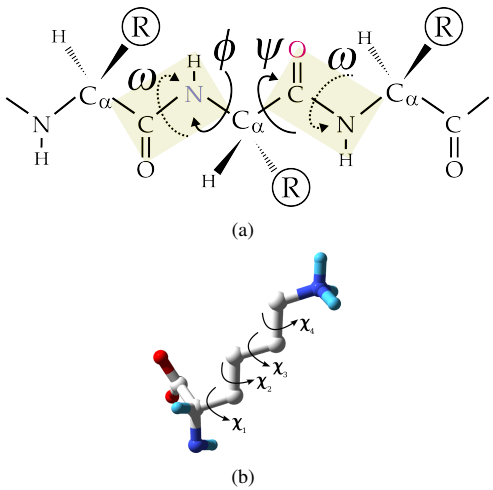


Fig. 1. (a) Main chain dihedral angles. "R" represents residues side chains. The plane of the peptide bond is the shaded square. (b) Side chain torsion angles of amino acid lysine.

The solutions were encoded as a chromosome with the same length of the amino-acids sequence. Each position in the chromosome consists of a data structure containing seven real values in the range $[-180°, 180°]$ (Figure 2). These are the backbone and side chain dihedral angles. Although the torsion angles of the peptide bond, ω, are explicit on the chromosome they do not change during the execution of the CSSGA and are kept in the *trans* (180°) configuration.

C. Fitness Function

The role of the fitness function is to evaluate the quality of a given candidate solution. The fitness is based on the energy

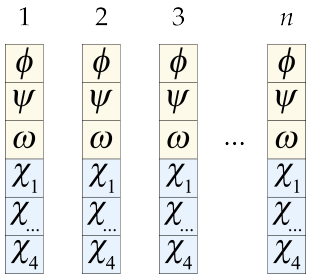


Fig. 2. Chromosome of length n encoding the backbone dihedral angles ϕ, ψ and ω and side-chain dihedral angles $\chi_1, \ldots, \chi_4$.

from the interaction between the atoms of the protein, calculated using the classical molecular force field GROMOS96 [27]–[29]. The force field terms modeling bond geometries are not used as they are invariant. The fitness function has the following form:

$$\begin{aligned} E_{total}(r_i) &= E_{torc} + E_{LJ} + E_{coul} + E_{solv} \\ E_{torc} &= \sum_{n}^{N_\phi} K_{\phi n}[1 + cos(n_n \phi_n - \delta_n)] \\ E_{LJ} &= \sum_{i \leq j}^{N_{\text{atoms}}} -\left(\frac{A_{ij}}{r_{ij}}\right)^6 + \left(\frac{B_{ij}}{r_{ij}}\right)^{12} \\ E_{coul} &= \sum_{i \leq j}^{N_{\text{atoms}}} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r(r_{ij}) r_{ij}} \end{aligned}$$

where r_{ij} is the distance between atoms i and j, A_{ij} and B_{ij} are Lennard-Jones parameters dependent of the atomic type, q_i and q_j are the atomic charges of atoms i and j, and ε_r is a distance dependent dielectric sigmoid function, which models the attenuation of the attraction between distant charges caused by water as a solvent, and preserves the strong attraction at shorter distances [30]. The parameter $K_{\phi n}$ is the energy constant associated with the torsion of a bond, ϕ_n is the torsion angle, n_n is the period and δ_n the phase angle. To increase the accuracy of the fitness function it would be interesting to model the interaction of the protein with the solvent. Since explicit modeling of the solvent molecules is infeasible, we used an approximate solvation model, developed in [31], which can, at least, further promote the burial of hydrophobic residues:

$$E_{solv} = \sum_i \Delta G_i^{ref} - \sum_{j<i} \left\{ \frac{2\Delta G_i^{free}}{4\pi^{\frac{3}{2}} \lambda_i d_{ij}^2} \exp\left(-r_{ij}^2\right) + \frac{2\Delta G_j^{free}}{4\pi^{\frac{3}{2}} \lambda_j d_{ij}^2} \exp\left(-r_{ij}^2\right) V_i \right\}$$

where d_{ij} is the sum of atomic van der Waals radii, λ is the correlation length, V is the atomic volume, and ΔG^{ref} and ΔG^{free} are empirically determined parameters associated with the atom type.

The resulting hyper-surface of a classical force field based fitness function is an extremely rough landscape with a large number of local optima, even when the only degrees of freedom considered are on the main chain dihedrals, ϕ and ψ [32].

D. Crowding-Based Steady-State Genetic Algorithm

The individuals in the initial population are generated slightly altering, by at most $20°$, the ϕ and ψ angles of an extended structure. Each of these structures has a $1/3$ probability of containing three random consecutive residues either in alpha-helix, beta-strand or random coil conformation. Then each individual has its energy evaluated by the fitness function and parents are randomly selected for the application of one genetic operator.

Six genetic operators are used:

1) 2-Point Crossover.
2) Multiple-point Crossover. The number of cut points is calculated as the closest integer to $l/10$, where l is the chromosome length [33].
3) Segment Mutation. This operator creates large variability. Three consecutive (ϕ, ψ) pairs have their values randomly changed to some other value in the interval $[-180.0°, 180.0°]$.
4) Incremental Mutation. A value from the interval $[-10.0°, 10.0°]$ is added to a single random ϕ or ψ.
5) Compensatory Mutation. With the goal of creating small chain movements, a randomly selected ϕ is incremented and the corresponding ψ is decremented by the same value.
6) Contiguous Swap. The values of ϕ and ψ of two contiguous positions are swapped.

Each time a ϕ or ψ value is changed, the corresponding rotamer for that residue is updated. The probability of application of each operator follows an adaptive scheme based on the quality of structures generated [34]. Initial probabilities are set to 16.67%, and are updated for each 10 new individuals. To avoid stagnation, the minimal allowed value is 2%.

After the creation of each new individual a crowding-based parental replacement procedure [35] is carried out. First, the structure encoded by the new individual is compared with all structures in the parental population. The fitness values are compared between the new individual and the most similar parent. If the new one has a better fitness it replaces that parent, otherwise it is discarded.

The structural similarity between individuals is given by the distance matrix error (DME) of the position of the alpha carbons of hydrophobic residues (as classified by [36]). For a structure with N hydrophobic amino acid residues this is calculated as:

$$\text{DME} = \sqrt{\sum_{i=1, j>i}^{N} \frac{(p_{ij} - q_{ij})^2}{\frac{N(N-1)}{2}}} \quad (1)$$

where p_{ij} and q_{ij} are the distances between the alpha carbons of residues i e j in the parental and offspring structures respectively.

E. Surrogate Modeling Approach

Surrogate modeling, also known as metamodeling, in general tries to capture in a simpler model the essential features of a simulation model (the original fitness function) [37] by approximating the input/output relation of the original model [38]–[40]. We used the similarity-based surrogate model described in [22] with adjustments required by its application in the context of a crowding-based steady-state GA with real encoding.

1) Similarity-based Surrogate Model: The model is based on the k-Nearest Neighbors method [20] where a database $\mathcal{D}$ storing information about η individuals, that have been evaluated by the original fitness function f, is maintained, i.e., $\mathcal{D} = \{[x^i, f(x^i)], i = 1, \ldots, \eta\}$. For a new individual x^h generated by a genetic operator, $\mathcal{D}$ is sorted by decreasing order of similarity to x^h, so that $\mathcal{I}$ is a list that stores individuals from the database most similar to x^h. The surrogate fitness value $\hat{f}(x^h)$ of x^h is:

$$\hat{f}(x^h) = \begin{cases} f(x^j) & \text{if } s(x^h, x^j) = 1 \text{ for some } j \in \{1, \ldots, \eta\} \\ \dfrac{\sum_{j=1}^{k} s(x^h, x^{\mathcal{I}_j})^u f(x^{\mathcal{I}_j})}{\sum_{j=1}^{k} s(x^h, x^{\mathcal{I}_j})^u} & \text{otherwise} \end{cases}$$

where $s(x^h, x^j)$ is a similarity measure between x^h and x^j, k is the number of nearest neighbors used to construct the model, and $u = 2$.

2) Similarity Measures: The foundation of the similarity-based surrogate model approach is the similarity principle, i.e., similar solutions have similar objective function values. To quantify how similar any two solutions are, two similarity measures were investigated. Based on the phenotypes, the first criterion is the same similarity measure used for the crowding procedure, i.e., the DME of the alpha carbons of hydrophobic residues. The second criterion is genotypic and is calculated as the Euclidean distance between the chromosomes, taking into account only the values of ϕ and ψ.

For any two solutions, their similarity is calculated as:

$$s(x^h, x^j) = 1 - d_H(x^h, x^j)$$

When using phenotypic similarity $d_H(x^h, x^j)$ is given by equation (1), and when using genotypic similarity it is calculated as the Euclidean distance between two chromosomes, where the difference δ between two angle values θ_1 and θ_2 is given by:

$$\delta = \begin{cases} |\theta_1 - \theta_2| & \text{if } |\theta_1 - \theta_2| < 180.0 \\ 360 - |\theta_1 - \theta_2| & \text{otherwise} \end{cases}$$

This same measure has been previously used to calculate the similarity between the angular portion of real-encoded chromosomes in a GA applied to the Flexible Ligand-Protein Docking Problem [41].

3) Surrogate Model Application and Management: When applied to the generational GA, a parameter p_{sm} controlled the fraction of the offspring population evaluated by the simulation model (the exact model). When applied to the CSSGA the parameter p_{sm} now controls the probability that a new individual will be evaluated by the simulation model.

The parameter p_{sm} also controls the fraction of the initial population that is evaluated by the simulation model. The remaining individuals are evaluated by the metamodel and the database $\mathcal{D}$, at the time of initialization, is composed by those individual evaluated by the simulation model.

To continuously refine the quality of the surrogate fitness approximation, the database $\mathcal{D}$ is updated to reflect the current general position the population occupies in the search space by using a model update procedure. Each new individual evaluated by the simulation model will replace the oldest individual in $\mathcal{D}$.

When comparing energy values, preference is given to the simulation model, i.e., if one of the two competing individuals' fitness values has been calculated by the simulation model that one is always selected. New structures evaluated by the metamodel can only enter the parental population when competing with a parent that also has been evaluated by the metamodel. This prompted some adaptations to use the surrogate model with the crowding methodology (section II-D), because the population can be quickly dominated by individuals evaluated by the exact objective function, thus making it impossible for new individuals evaluated by the metamodel to enter the population. To remedy this situation crowding is performed only amongst individuals evaluated by their respective model, e.g., when a new individual is about to enter the population, its similarity is only calculated against individuals evaluated by the same objective function, i.e., exact or surrogate. The effect is that the population always has a fraction $(1 - p_{sm})$ of it evaluated by the metamodel.

III. Results

A. Test Set

In order to assess the effects of the surrogate model, a set of six small proteins with known structures, shown in Table I, was used.

TABLE I
DESCRIPTION OF THE TEST PROTEINS.

PDB ID [42]	Sequence length	Atoms*	Class
23ala	23	141	alpha
1e0n	27	299	beta
1amb	28	305	alpha
1vii	36	389	alpha
1l2y	37	355	alpha
1e0l	37	410	beta

*The number of atoms is for the structure modeled under the GROMOS96 force field.

B. Parameters

The parameters for the GA used during the tests were: population size of 200, chromosome lengths according to Table I, operators described in section II-D with adaptive probabilities, a maximum number of $200,000$ exact function evaluations ($N_{f,max}$), and 30 independent runs per sequence.

The surrogate model was constructed based on a database of $\eta = 2000$ solutions.

The quality of the generated structures was evaluated by their energy values and it is considered that lower energies correspond to better structures. Only energies were used to measure the effects of the metamodel introduction, because they are the direct product of optimization by the GA. Structural comparisons between the generated models and the experimentally determined structure, e.g. Root Mean Square Deviation (RMSD) of backbone atoms, would not reveal these effects as efficiently because there is no direct correlation between the RMSD and the energy, that is, there could be structures very similar to the native structure, yet with very high energies. Statistics were calculated from the best individual evaluated by the exact model on the populations at the end of the runs.

C. Impact of the Surrogate Modeling Approach

To measure the impact of the surrogate model different values of the fraction p_{sm} were used, that is, $p_{sm} \in [0.1, 1.00]$, e.g., for $p_{sm} = 0.2$ and $N_{f,max} = 200,000$ up to $160,000$ extra surrogate evaluations are performed. When $p_{sm} = 1.0$ the standard CSSGA is recovered, i.e., only the exact fitness function is used. This is useful to assess the effects of increasing the number of meta-evaluations performed compared to the standard CSSGA.

Figure 3 summarizes the results obtained from the best solutions at the end of each of the 30 runs for all sequences. It can be observed that the introduction of the surrogate model in the evolutionary model cycle can improve the quality of final solutions, with optimal values for p_{sm} between 0.5 and 0.9. One common behavior observed from all sequences is that as p_{sm} decreases, the quality of solutions also decreases. This is expected because since that parameter directly controls the fraction of individuals in the population evaluated by the simulation model it affects the update of the database $\mathcal{D}$. The number of neighbors k used during these runs was 3.

When comparing the results of the two similarity measures, significant differences in the performance cannot be observed when using the genotypic or the phenotypic distance to perform the surrogate fitness evaluation. Except that, for low values of p_{sm} (≤ 0.3), the use of the phenotypic criterion tends to yield worse energies.

The results for 23ala can be explained by using some insight about the protein. Polyalanine is a small peptide whose native structure is an alpha-helix. These peptides have been previously studied by various optimization techniques, including methodologies with multiple minima features [43], [44]. An important feature of polyalanines is that they have less complexity, when compared with other sequences, because there are no movements in the side chains (alanine has no torsion angles χ_i in its side chain). Additionally, an alanine residue has its side chain modeled as a united atom CH_3 under the GROMOS96 force field, and the CH_3 has no polarity or charge, therefore it does not contribute to the E_{coul} potential. It can be observed that the allowed number of exact function evaluations is sufficient to reach the native structure for that sequence, which exhibits fitness scores of $\sim$ 105Kcal/mol and thus the use of the surrogate model presents very little changes on the performance, except when using very low values for p_{sm}, see Figure 3(a).

D. Analysis of the number of neighbors

The effects of using different values for k during calculation of the surrogate models were investigated for a value of $p_{sm} = 0.7$. This value was chosen by repeating the experiments described on section III-B with 1 and 7 neighbors and it is centered on the optimal range found (0.6 to 0.9). To further identify an optimal number of neighbors, the results of 30 runs were compared for each sequence using the same parameters described on section III-B and $p_{sm} = 0.7$ for 1, 3, 5, 7, 9 and 15 neighbors.

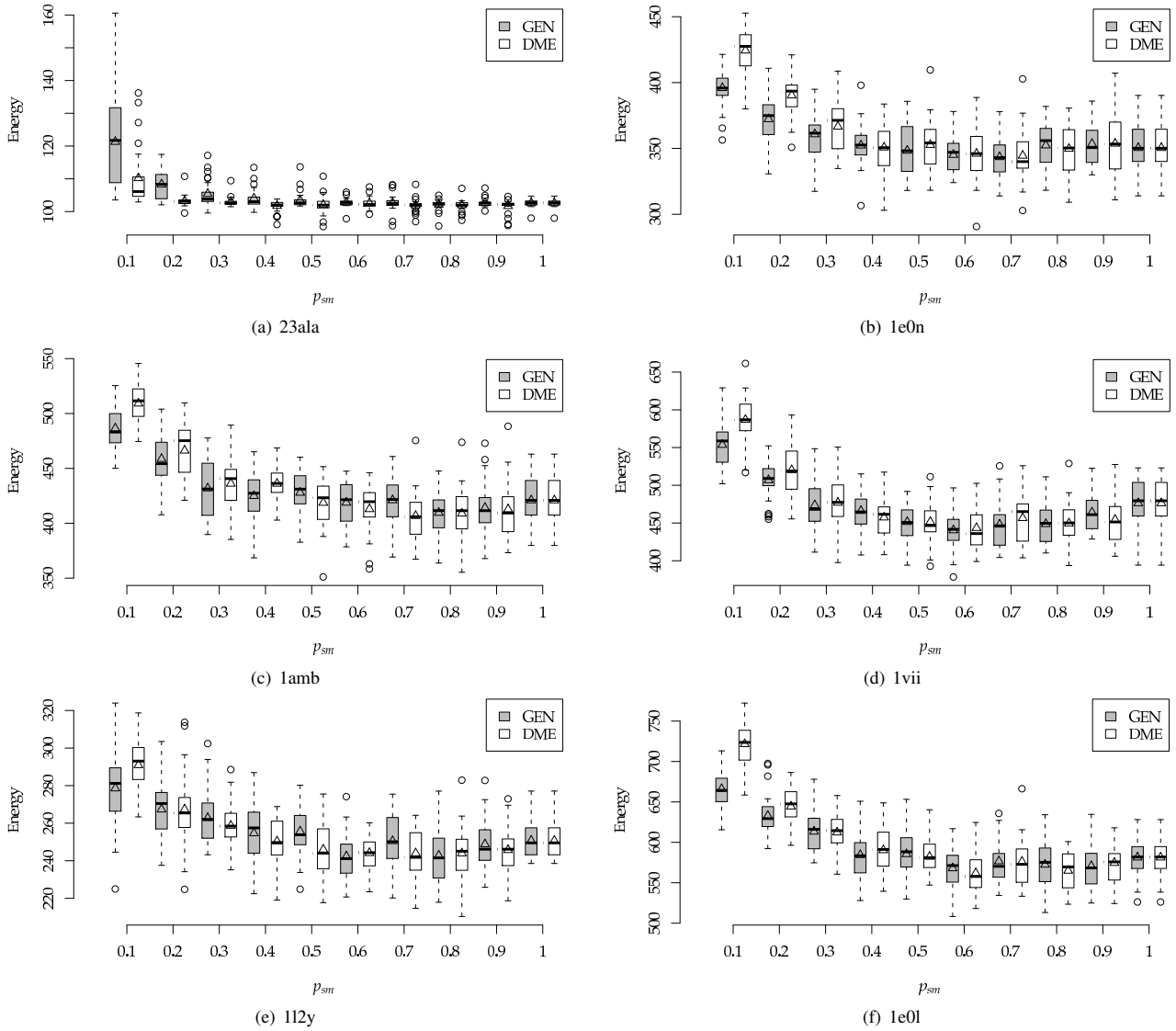
Figure 4 summarizes the results obtained from the best solutions at the end of each of the 30 runs for all sequences. There is no significant difference among the results obtained using 1, 3, 5 or 7 neighbors, as was confirmed by an ANOVA test. However, the quartiles from the boxplots showed a tendency to represent worse energies as k increases for both distance criteria (Figure 4).

IV. Conclusions

In this paper a similarity-based surrogate-model approach was tested with a real-encoded crowding-based steady-state genetic algorithm applied to the *ab initio* Protein Structure Prediction problem. Using two measures of similarity, the effects of increasing the number of surrogate fitness evaluations was investigated and compared to the standard CSSGA performing only the same number of exact fitness evaluations.

The results show that the use of surrogate fitness evaluations, particularly adopting a phenotypic distance criterion between solutions, can improve the quality of the solutions. However the presented technique showed limitations. For instance, it is important to maintain a relatively high value of p_{sm} as this not only controls the probability of individuals being evaluated by the exact fitness function, but also reflects on the update ratio of the database $\mathcal{D}$ and this can affect the accuracy of the approximations.

It is important to note that the total computation cost increases as more surrogate evaluations are performed, even if they represent only a fraction of the computational cost of the exact function. Also, the DME calculation has a slightly higher computational cost (less than 1%) than the simple chromosome Euclidean distance calculation. The cost of a single exact function evaluation is 3.5 times higher than a surrogate evaluation, for the structures with fewer atoms (23ala), and about 31 times higher for the larger structures (1e0l). This difference will be even greater as the structures' sizes increase. The metamodel approach was implemented into our protein structure prediction program, developed in C++, and is not yet fully optimized. Preliminary measurements using $k = 3$ for 23ala and 1e0l yielded an average wall clock per run of 3.5min and 26min using



(a) 23ala (b) 1e0n

(c) 1amb (d) 1vii

(e) 1l2y (f) 1e0l

Fig. 3. Results from 30 runs for each sequence, with $k = 3$. Grey box plots are results obtained using the genotypic similarity measure (Euclidean) while the white ones are using phenotypical (DME). The triangles are the mean results. Statistics are calculated from the best individual at the end of each run. Energies in the y-axis are in Kcal/mol (lower is better). The value of p_{sm} shown in the x-axis is the probability of a new individual being evaluated by the simulation model.

$p_{sm} = 1.0$ and 4min and 28m using $p_{sm} = 0.7$, respectively, on an Intel(R) Core(TM)2 Quad CPU Q9550 @ 2.83GHz.

Protein structure prediction experiments can benefit from the use of k-NN surrogate model with values for p_{sm} from 0.6 to 0.9 and $k = 3$, as the results showed that this configuration improves the overall quality of the solutions for the four longest test sequences (Figure 3), with minimal increase of the total computational cost over that of a standard CSSGA.

In theory, to use more information about the protein structure itself should provide better approximations for the k-NN surrogate model. Possibly more detailed structural comparison methods, such as the root mean squared deviation (RMSD) of all atoms will provide even better approximations. In future work we will be investigating these more detailed similarity measures as well as other metamodeling techniques.

ACKNOWLEDGMENT

This work was supported by FAPERJ (grants E-26/171.401/01, E-26/170.648, E-26/102.443/2009 and E-26/102.825/2008) and CNPq (grants 151594/2008-2 and 311651/2006-2).

REFERENCES

[1] H. Höltje, W. Sippl, D. Rognan, and G. Folkers, *Molecular modeling: basic principles and applications*. Vch Verlagsgesellschaft Mbh, 2008.

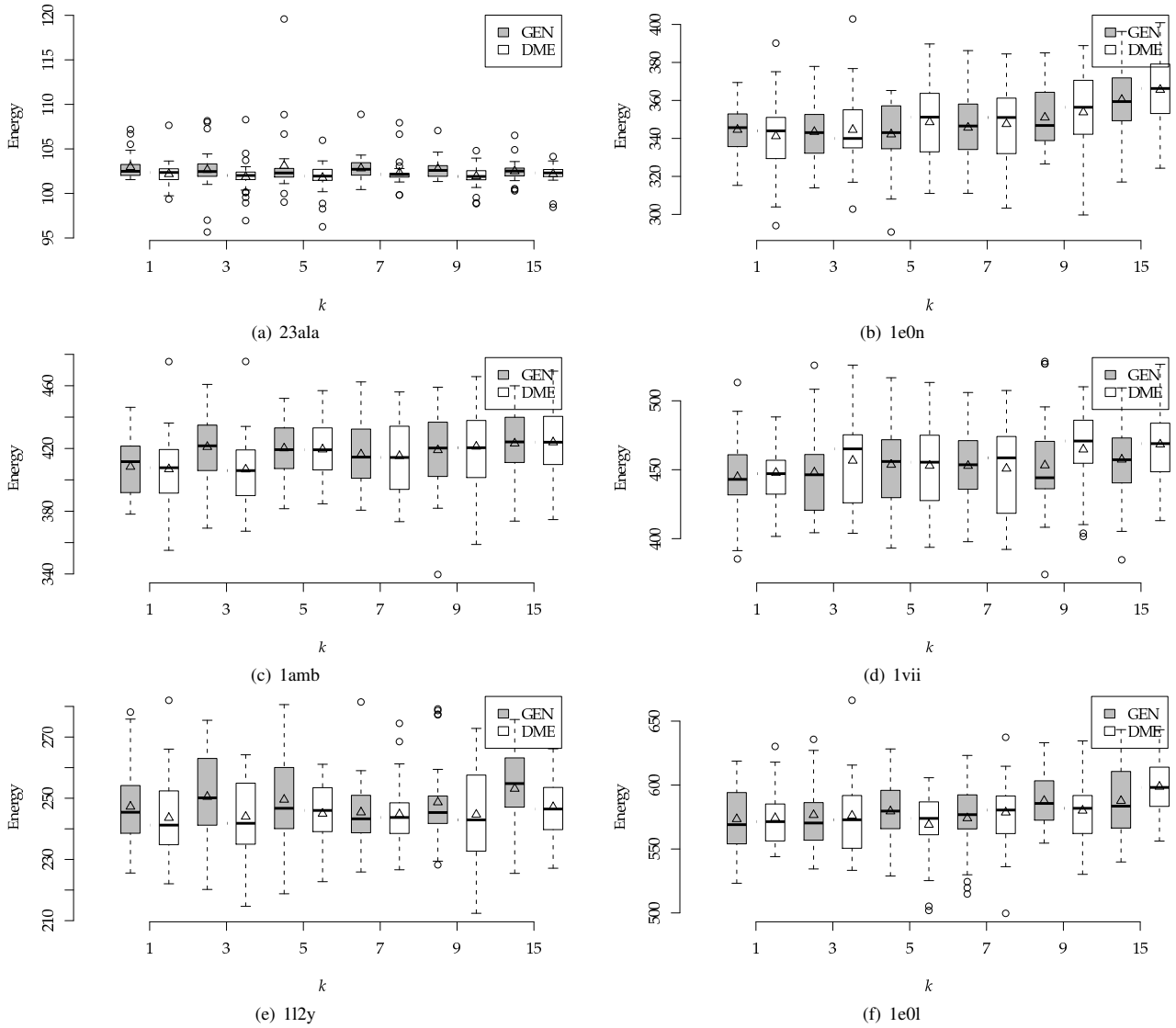


Fig. 4. Results from 30 runs for each sequence using $p_{sm} = 0.7$. Grey box plots are results obtained using genotypic similarity measure (Euclidean) while the white ones are using phenotypical (DME). The triangles are the mean results. Statistics are calculated from the best individual at the end of each run. Energies in the y-axis are in Kcal/mol (lower is better). The value of k shown in the x-axis is the number of neighbors used to calculate the surrogate model.

[2] M. Ben-David, O. Noivirt-Birk, A. Paz, J. Prilusky, J. Sussman, Y. Levy, and E. Pearl, "Assessment of CASP8 structure predictions for template free targets," *Proteins*, vol. 77, no. Suppl 9, pp. 50–65, 2009.

[3] D. Rthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, and D. Baker, "Kemp elimination catalysts by computational enzyme design.," *Nature*, vol. 453, pp. 190–195, May 2008.

[4] I. W. Davis and D. Baker, "RosettaLigand docking with full ligand and receptor flexibility.," *J Mol Biol*, vol. 385, pp. 381–392, Jan 2009.

[5] B. Qian, A. Ortiz, and D. Baker, "Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation," *Proceedings of the National Academy of Sciences*, vol. 101, no. 43, p. 15346, 2004.

[6] Y. Shen, R. Vernon, D. Baker, and A. Bax, "*De novo* protein structure generation from incomplete chemical shift assignments," *Journal of Biomolecular NMR*, vol. 43, no. 2, pp. 63–78, 2009.

[7] C. B. Anfinsen, "Principles that govern the folding of proteins," *Science*, vol. 181, p. 187, 1973.

[8] K. M. S. Misura, D. Chivian, C. A. Rohl, D. E. Kim, and D. Baker, "Physically realistic homology models built with Rosetta can be more accurate than their templates," *Proceedings of the National Academy of Sciences*, vol. 103, no. 14, pp. 5361–5366, 2006.

[9] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. Strauss, and D. Baker, "Rosetta in CASP4: progress in ab initio protein structure prediction.," *Proteins*, vol. Suppl 5, pp. 119–126, 2001.

[10] Y. Zhang, "Template-based modeling and free modeling by I-TASSER in CASP7.," *Proteins*, vol. 69, pp. 108–117, Sep 2007.

[11] J. H. Holland, *Adaptation in Natural and Artificial Systems*, vol. 183. University of Michigan Press, 1975.

[12] R. Myers, D. Montgomery, and C. Anderson-Cook, *Response surface methodology: process and product optimization using designed experiments*. Wiley, 2009.

[13] A. Mullur and A. Messac, "Metamodeling using extended radial basis functions: a comparative approach," *Engineering with Computers*,

vol. 21, no. 3, pp. 203–217, 2006.
[14] M. Hussain, R. Barton, and S. Joshi, "Metamodeling: Radial basis functions, versus polynomials," *European Journal of Operational Research*, vol. 138, no. 1, pp. 142–154, 2002.
[15] S. Ferrari and R. Stengel, "Smooth function approximation using neural networks," *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 24–38, 2005.
[16] M. Emmerich, K. Giannakoglou, and B. Naujoks, "Single-and multiobjective evolutionary optimization assisted by gaussian random field metamodels," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 4, pp. 421–439, 2006.
[17] V. Kecman, *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. The MIT press, 2001.
[18] J. Branke and C. Schmidt, "Faster convergence by means of fitness estimation," *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, vol. 9, no. 1, pp. 13–20, 2005.
[19] Y. Jin, "A comprehensive survey of fitness approximation in evolutionary computation," *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, vol. 9, no. 1, pp. 3–12, 2005.
[20] N. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
[21] F. Custódio, H. Barbosa, and L. Dardenne, "Genetic Algorithm for Finding Multiple Low Energy Conformations of Poly Alanine Sequences Under an Atomistic Protein Model," *Lecture Notes in Computer Science*, vol. 4643, p. 163, 2007.
[22] L. G. Fonseca, H. J. C. Barbosa, and A. C. C. Lemonge, "A similarity-based surrogate model for expensive evolutionary optimization with fixed budget of simulations," in *CEC'09: Proceedings of the Eleventh conference on Congress on Evolutionary Computation*, pp. 867–874, Institute of Electrical and Electronics Engineers Inc., The, 2009.
[23] J. Ponder and D. Case, "Force fields for protein simulations," *Advances in protein chemistry*, vol. 66, pp. 27–86, 2003.
[24] S. Schulze-Kremer, *Genetic Algorithms and Protein Folding Protein Structure Prediction Methods and Protocols*, vol. 143, ch. 9, pp. 175–221. Humana Press Inc., 2000.
[25] F. Jensen, *Introduction to computational chemistry*. Wiley New York, 1999.
[26] R. L. Dunbrack, "Rotamer libraries in the 21st century.," *Curr Opin Struct Biol*, vol. 12, pp. 431–440, Aug 2002.
[27] W. F. van Gunsteren and H. J. C. Berendsen, "Groningen molecular simulation (GROMOS) library manual," *Biomos, Groningen*, 1987.
[28] L. J. Smith, A. E. Mark, C. M. Dobson, and W. F. van Gunsteren, "Comparison of MD simulations and NMR experiments for hen lysozyme. Analysis of local fluctuations, cooperative motions, and global changes.," *Biochemistry*, vol. 34, pp. 10918–10931, Aug 1995.
[29] L. D. Schuler, X. Daura, and W. F. van Gunsteren, "An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase," *Journal of Computational Chemistry*, vol. 22, no. 11, pp. 1205–1218, 2001.
[30] N. Arora and B. Jayaram, "Strength of hydrogen bonds in alpha-helices," *Journal of Computational Chemistry*, vol. 18, no. 9, pp. 1245–1252, 1997.
[31] T. Lazaridis and M. Karplus, "Effective energy function for proteins in solution.," *Proteins*, vol. 35, pp. 133–152, May 1999.
[32] A.-A. Tantar, N. Melab, and E.-G. Talbi, "A comparative study of parallel metaheuristics for protein structure prediction on the computational grid," *Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International*, pp. 1–10, 26-30 March 2007.
[33] F. L. Custódio, H. C. J. C. Barbosa, and L. E. Dardenne, "Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm," *Genetics and Molecular Biology*, vol. 27, pp. 611 – 615, 00 2004.
[34] L. Davis, *Handbook of genetic algorithms*. Boston: London International Thomson Computer Press, 1996.
[35] K. A. DeJong, *Analysis of the Behavior of a Class of Genetic Adaptive Systems*. PhD thesis, Computer and Communication Sciences, University of Michigan, 1975.
[36] I. Callebaut, G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat, and J. P. Mornon, "Deciphering protein sequence information through hydrophobic cluster analysis (hca): current status and perspectives.," *Cell Mol Life Sci*, vol. 53, pp. 621–645, Aug 1997.
[37] W. Hendrickx, D. Gorissen, and T. Dhaene, "Grid enabled sequential design and adaptive metamodeling," in *WSC '06: Proceedings of the 38th Conference on Winter Simulation*, pp. 872–881, Winter Simulation Conference, 2006.
[38] R. R. Barton, "Metamodels for simulation input-output relations," in *WSC '92: Proceedings of the 24th Conference on Winter Simulation*, (New York, NY, USA), pp. 289–299, ACM, 1992.
[39] J. Kleijnen and R. Sargent, "A methodology for fitting and validating metamodels in simulation," *European Journal of Operational Research*, vol. 120, no. 1, pp. 14–29, 2000.
[40] D. Fonseca, D. Navaresse, and G. Moynihan, "Simulation metamodeling through artificial neural networks," *Engineering Applications of Artificial Intelligence*, vol. 16, no. 3, pp. 177–183, 2003.
[41] C. S. de Magalhães, H. J. C. Barbosa, and L. E. Dardenne, "Selection-insertion schemes in genetic algorithms for the flexible ligand docking problem," *Genetic and Evolutionary Computation - GECCO 2004*, pp. 368–379, 2004.
[42] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Res*, vol. 28, pp. 235–242, Jan 2000.
[43] E. Clementi and S. Chin, *Biological and artificial intelligence systems*. Leiden: ESCOM, 1988.
[44] H. A. Scheraga, J. Lee, J. Pillardy, Y. J. Ye, A. Liwo, and D. Ripoll, "Surmounting the multiple-minima problem in protein folding," *Journal of Global Optimization*, vol. 15, no. 3, pp. 235–260, 1999.