Combination of Differential Evolution and Fragment-based Replacements for Protein Structure Prediction

Daniel Varela
Department of Computer Science
University of A Coruña
Campus de Elviña s/n, 15071 A Coruña (Spain)
daniel.varela@udc.es

José Santos
Department of Computer Science
University of A Coruña
Campus de Elviña s/n, 15071 A Coruña (Spain)
santos@udc.es

ABSTRACT

In this work Differential Evolution (DE) was combined with fragment replacement for improving the search of protein structure conformations with minimum energy. The Rosetta environment was used, employing some of its phases for the ab initio prediction in the initialization of the genetic population, as well as its fragment-assembly technique. DE provides a global search in the multimodal energy landscape whereas fragment replacement based on the Monte-Carlo procedure provides a useful search of short protein conformations that accelerates the DE search. Initial results with proteins from PDB with a comparison with previous works are provided.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and medical sciences

Keywords

Protein structure prediction, Evolutionary computing, Differential evolution

1. INTRODUCTION

Knowledge of the three-dimensional structure of a protein will help us to understand the questions about protein function and drug design. A large number of protein sequences was obtained thanks to the success of the genome sequence projects, but it is not possible to determinate all the protein structures due to the technical difficulties and the time cost, therefore it is necessary to use the computational prediction.

The discovered conformations can be used to create templates and help to predict the conformation of similar proteins, but if there is not a template available for the protein to solve, we have to determine the conformation from scratch. This procedure is called *ab initio*, which tries to determine the final structure only from the information of the protein primary sequence. It is based on Anfinsen's dogma [1], which postulates that the native structure is determined only by the protein's amino acid sequence, and the thermodynamic hypothesis that states that the biologically native fold is a free energy minimum. This is a challenge for computational biology and it can be understood as an optimization problem.

Typically, ab initio protocols are split in two stages [6]. The first stage consists of a conformational search guided by an energy function with the objective of generating a number of possible low-energy conformations (decoys). The second stage selects a subset of these decoys that are relevant for the native state of the protein and performs a high-resolution structural refinement which is a computational intensive process. Rosetta uses a physics and knowledge-based energy function [6]. Knowledge-based potential [15] refers to the empirical energy terms derived from the statistics of the solved structures deposited in PDB [9]. This function is used during the conformational search, combined by assembling small protein fragments [14] (3-mers and 9-mers) taken from the PDB library to narrow the conformational search as it is explained further. Physics-based energy function [4] contains terms associated with bond lengths, angles, torsion angles, Van der Waals and electrostatic interactions.

In the case of Rosetta, ab initio implementation [11] starts with a stage that uses off-lattice representation, also known as coarse-grained representation, which consist of only the ϕ, ψ and ω dihedral angles with the side chains described by a centroid located at the center of mass. To search through the conformational space, many rounds of Metropolis Monte Carlo (MMC) are performed. This algorithm consists of generating a trial conformation by doing fragment replacements. The Metropolis criterion is used to determine the acceptance of this move by calculating the energy difference following a Boltzmann energy distribution for a given temperature (a fixed temperature of 2 is used with a possibility of re-heating). In the final optional stage, known as Relax, a high time-consuming refinement using an all-atom representation is performed to obtain a near native conformation.

There are many works using evolutionary computing methods and other natural computing approaches to the protein structure prediction problem, especially using lattice models for representing the protein conformations [17][13]. Nevertheless, there are few works using evolutionary computing with off-lattice models and the Rosetta environment, such as the works by Olson et al. [7][8]. In this work we used the Rosetta environment and tested the use of Differential Evolution (DE) [10], a robust method in many optimization problems, adapted to our application and hybridized with the Rosetta fragment replacements. The aim is to im-

GECCO '15, July 11 - 15, 2015, Madrid, Spain

DOI: http://dx.doi.org/10.1145/2739482.2768437

Algorithm 2.1: HYBRID DIFFERENTIAL EVOLUTION()

```
for each Individual ∈ Population
  do {Individual ← INITIALIZERANDOMPOSITIONS()
repeat
 // DE Phase
 for each Individual  x ∈ Population
      do { x1, x2, x3 ← GETRANDOMINDIVIDUAL(Population)
           // must be distinct from each other and x
           R ← GETRANDOM(1, n)  // the highest possible
           // value n is the dimensionality of the problem to be
           // optimized
           for each i ∈ 1 : n
           // Compute individual's potentially new position
           // y = [y1, ..., yn]
                do { ri ← GETRANDOM(0, 1)// uniformly in
                     // open range (0,1)
                     if ((i = R) || (ri < CR))
                          yi = x1i + F(x2i − x3i)
                       else yi = xi
           if (f(y) ≤ f(x))   x = y
           // replace x with y in Population
 // Phase of fragment replacements in the population
 for each Individual  x ∈ Population
      do { for each j ∈ 1 : number of amino acids
           // j refers to the amino acid position
           rj ← GETRANDOM(0, 1)
           rmode ← GETRANDOM(0, 1)
           if (rj < PR) // PR - probability of replacement
           if (rmode < 0.5)
           REPLACEMENT ATTEMPT(j, 3mer fragment)
           // Replacement in position j with a 3-mer fragment
           // The replacement is applied only if it improves energy
             else REPLACEMENT ATTEMPT(j, 9mer fragment)
           // Replacement in position j with a 9-mer fragment
           // The replacement is applied only if it improves energy
until TERMINATIONCRITERION()
return (GETLOWESTFITNESS(Population))
// return candidate solution
```

prove the sampling of promising areas in the energy landscape of protein conformations, integrating, for a more efficient search, the advantages of DE as a global search method with the search of short conformations provided by fragment replacements.

2. METHODS

2.1 Differential Evolution

Differential Evolution [10] is a population-based search method. DE creates new candidate solutions by combining existing ones according to a simple formula of vector crossover and mutation, and then keeping whichever candidate solution has the best score or fitness on the optimization problem at hand. The central idea of the algorithm is the use of difference vectors for generating perturbations in a population of vectors. This algorithm is specially suited for optimization problems where possible solutions are defined by a real-valued vector. The basic DE algorithm is summarized in the pseudo-code of Algorithm 2.1 (DE phase).

Differential Evolution needs a reduced number of parameters to define its implementation. The parameters are F or differential weight and CR or crossover probability. The weight factor F (usually in $[0, 2]$) is applied over the vector resulting from the difference between pairs of vectors (x_2 and x_3). CR is the probability of crossing over a given vector of the population (target vector x) and a "donor" vector created from the weighted difference of two vectors ($x_1 + F(x_2 - x_3)$) [2]. The "binomial" crossover (specified in the pseudo-code of Algorithm 2.1), for defining the value of the "trial" vector (y) in each vector component or position i [2], was used. Finally, the index R guarantees that at least one of the parameters (genes) will be changed in the generation of the trial solution.

Finally, the selection operator maintains constant the population size. The fitness of the trial vector ($f(y)$) and the target vector ($f(x)$) are compared to determine which one survives for the next generation: If the new trial vector yields an equal or lower (better) value of the objective function, it replaces the corresponding target vector in the next generation; otherwise the target vector is retained [2]. Thus, the fitness of the best solution of the population is improved or remains the same through generations.

As Feoktistov [3] indicates, the fundamental idea of the algorithm is to adapt the step length ($F(x_2 - x_3)$) intrinsically along the evolutionary process. At the beginning of generations the step length is large, because individuals are far away from each other. As the evolution goes on, the population converges and the step length becomes smaller and smaller, providing this way an automatic balance in the search.

2.2 Protein conformation encoding

The coarse-grained representation of Rosetta [12] was used. This centroid mode considers the location of the main backbone atoms, whereas each side chain is represented by a united pseudo-atom located at the side-chain center of mass.

Each protein conformation is encoded with the three dihedral angles, ϕ, ψ and ω, for each amino acid. The application of forward kinematics to this angular representation obtains the spatial information of the protein conformation. DE individuals code the dihedral angles in the range [-1,1], which are decoded to the interval [-180,180] (degrees). The angles ϕ and ψ are evolved with DE whereas the third dihedral angle, ω, is not evolved. This last angle is only changed by fragment replacements. The reason is that this angle can only have two configurations of $180°$ or $-180°$.

2.3 Protein conformational energy

We used the Rosetta energy functions to calculate the free energy of each conformation in the population. The Rosetta energy score of a protein is a linear combination of weighted terms that models molecular forces that act on and between all atoms in that conformation. There are energy terms such as solvation and electrostatics effects, repulsion, hydrogen bonding, and secondary structure scores such as strand pairing and helix-strand packing. Steric overlap of backbone atoms and side-chain centroids is penalized, but favorable Van der Walls interactions are modeled only by rewarding globally compact structures [11].

The Rosetta score function which takes into account all energy components, called *score3*, corresponds to the full coarse-grained energy function and it is used as fitness function in the Differential Evolution algorithm. Nevertheless, Rosetta changes the weight set depending on the stage of its ab initio protocol. For instance, *score0* is used during

the first stage, which considers only a steric repulsion term, while in the second stage Rosetta uses a more complex score function, *score1*, incorporating energy terms to score secondary structure interactions.

2.4 Population initialization

The first stages of the Rosetta ab initio procedure were used for initializing the individuals of the population. As commented previously, Rosetta uses a Metropolis Monte Carlo method, which is divided in four stages. Along these stages, Rosetta uses the low-resolution description (coarse-grained representation) for the protein and a fragment insertion technique to generate new decoys.

The first Rosetta stage begins with a fully extended chain and inserts 9-mer fragments until all the backbone angles were modified at least once and with a maximum of 2000 cycles. During this stage, the energy function only considers the steric-clash term to prevent overlapping between backbone atoms and side-chain centroids.

The second Rosetta stage employs 9-mer fragment insertions during 2000 cycles, but the scoring function adds terms such as hydrophobic burial and specific pair iterations, as well as secondary structure scores. In the proposed methodology with DE, only 200 cycles are performed per individual. Moreover, we used 3-mer and 9-mer fragment insertions in this stage.

Therefore, each individual of the genetic population is the result of the application of these two first phases of Rosetta, generating different protein conformations that define the initial population. The idea is to use these fast initial stages of Rosetta to sample the conformational space to obtain suitable initial conformations across the search space. On the contrary, using individuals with rotating angles defined by values sampled uniformly at random, the evolutionary algorithm should discover, first, appropriate combinations of angles according, for example, to the secondary structure elements and Ramachandran intervals.

2.5 Hybrid DE using fragment replacements

After the DE phase, all individuals of the genetic population are "mutated" using the Rosetta fragment replacements. Rosetta uses a library of short peptide fragments (typically 3 and 9 residue long) as a Monte Carlo moves set. Each fragment is defined by the three backbone dihedral angles per residue. Each time a fragment is inserted in an encoded protein of the genetic population, a number of subsequent angles are affected in the encoded conformation.

On the contrary to the first Rosetta phases used in the initialization of the population, these replacements are only accepted if they improve (decrease) the energy of the resulting protein conformation after the replacement. Obviously, the use of the Metropolis Monte Carlo method employed in Rosetta can increase the diversity in the genetic population, since worse replacements could also be accepted (depending on the temperature parameter used in the Metropolis criterion). Nevertheless, the DE genetic operators (crossover and the generation of the mutant vector) were sufficient to maintain an appropriate diversity in the population.

In the hybridization with DE, for each individual a probability to decide if a replacement is checked in each amino acid position j (parameter PR in Algorithm 2.1) was used. Moreover, two fragment libraries with lengths 3 and 9 residues were employed, with the same probability to use a 3-mer or a 9-mer fragment replacement, since the optimal fragment length varies for different proteins [5]. It should be noted that these replacements are applied after the DE phase, as schematized in Algorithm 2.1.

There would be another possibility of hybridization, applying the replacements after the trial vector is calculated. However, this alternative would eliminate the main idea and advantage of DE with the calculation of the donor or mutant vector by means of the difference vectors. Therefore, the selected alternative where the replacements are applied out of the DE main procedure was used.

3. RESULTS

In Differential Evolution a standard value for the CR parameter was used ($CR = 0.95$), in the interval suggested in [16] ($CR \in [0.8, 1.0]$), whereas a low value for the F parameter ($F = 0.03$) was employed. The reason of the low value is because a very small change in the angle values can imply a high change in the conformational energy, so DE must change the angles of the base vector (x_1) with low values when calculating the donor and trial vectors (Algorithm 2.1), taking into account that these changes can affect simultaneously to a large number of amino acid angles. In the case of the population size, a fixed number of 1000 individuals in the different runs was used.

The DE variant $DE/rand/1/bin$ (where 1 denotes the number of differences involved in the construction of the mutant or donor vector and *bin* denotes the crossover type) was employed, variant which chooses the base vector x_1 randomly, providing low selective pressure in the runs.

For the initial population, as previously explained, the first two phases of Rosetta were applied to define the initial individuals. In the first stage we maintained the maximum of 2000 cycles used by Rosetta when trying that all the backbone angles are modified at least once. In the second stage, on the contrary to the Rosetta fragment insertions during 2000 cycles, only 200 cycles per individual were used, which are sufficient to the objective of an initial population dispersed across the conformational space and with the angles initialized within suitable interval values. Hence, the angles in the initial individuals are in typical ranges of the corresponding secondary structure (predicted by Rosetta or specified in the fragments replaced).

Each run of the hybrid DE is applied for a fixed budget of 10,000,000 energy function evaluations, the same used in [7]. As indicated, after the DE phase, for each individual, we used a probability ($PR = 0.01$) to decide, in each amino acid position, if a replacement (3-mer or 9-mer fragment) is checked for energy improvement. Obviously, each replacement test implies an energy evaluation, plus the additional energy evaluation per individual in the DE phase.

The authors in [7] also worked with the Rosetta coarse-grained representation and applied a hybrid evolutionary algorithm (EA) combined with a local search to map each child conformation to a nearby local minimum (using fragment replacements). Their EA used the classical operators of crossover (1-point and 2-point crossover as well as a homologous crossover) and a mutation operator implemented also with the fragment replacements. Olson et al. [8] also experimented with the use of multi-objective optimization for the same problem but they reported results using the *score 4* Rosetta function (used to compare the obtained decoys before the last Rosetta refinement).

	PDB Id	Size	Native Fold Topology	Lowest Energy Value in [7]	Lowest Energy Value	RMSD
1	1dtdB	61	α/β	-38.8 (-19.1)	-42.19 (-19.62)	9.24 (10.54)
2	1c8cA	64	α/β	-68.3 (-45.9)	–92.07 (-75.37)	4.88 (8.37)
3	1sap	66	α/β	-86.2 (-58.6)	-106.98 (-79.75)	3.23 (8.11)
4	1hz6A	67	α/β	-99.8 (-64.7)	-72.65 (-48.39)	12.80 (13.23)
5	1wapA	68	β	-93.3 (-60.8)	-113.50 (-66.38)	11.55 (12.50)
6	1ail	70	α	-27.3 (-13.6)	-28.81 (-18.99)	12.27 (10.38)

Table 1: Energy values (*score3*) obtained with different PDB proteins. In the first columns: PDB Id of the native structure, number of amino acids, and native fold topology for each target protein. Column 5 shows the minimum lowest energy over 30 different runs as well as the average lowest energy in the 30 runs (in parentheses) reported by Olson et al. [7] using a hybrid EA. Column 6 shows the best energy value in 10 independent runs and the average lowest value (in parentheses) in the 10 runs using the hybrid DE. Column 7 shows the RMSD of the best conformation found as well as the average RMSD in the best conformations in the 10 independent runs.

In Table 1 we included the comparison of the results using the hybrid DE with the results of the hybrid EA employed by Olson et al. [7] using 6 PDB proteins. Table 1 includes the best energy values (*score3*) obtained in 10 independent runs with the hybrid DE (last column). The value in parentheses corresponds to the average of the best values in those independent runs. For PDB proteins *1dtdB*, *1sap* and *1wapA*, Olson et al. [7] obtained the best results (regarding the average best values in 30 runs) when they used 1-point crossover and mutation as genetic operators in their hybridized EA. For the best results of protein *1ail* they used 2-point crossover and mutation whereas with proteins *1c8cA* and *1hz6A* they used a homologous crossover and mutation for obtaining the best results [7].

Nevertheless, using only a genetic operator in DE that combines mutation and crossover, it allows the hybrid DE solution to obtain best energy values (the best value and average best values) with respect to the hybrid EA [7] (except in protein *1hz6A* with the limited number of energy evaluations). The reason is in the simple formula of perturbation of the genetic population used by DE (difference of vectors), which provides an appropriate search in the vicinity of the encoded protein conformations, whereas the replacement procedure searches for suitable angle combinations for the refinement with DE. On the contrary, the RMSD values of the found best conformations present higher values with respect to the values in [7]. However, it must be taken into account the inaccuracies in the Rosetta energy function as well as that Olson et al. [7] reported the best RMSD values found in the course of the genetic search, whereas the reported values in Table 1 correspond to the values of the final best conformations found by the hybrid DE algorithm.

Finally, the main drawback of the hybrid DE solution is the difficult control of premature convergence, even with the low selective pressure applied, therefore self-adaptive schemes of DE parameters [2] will be tested.

4. ACKNOWLEDGMENTS

This work was funded by the Ministry of Economy and Competitiveness of Spain (project TIN2013-40981-R).

5. REFERENCES

[1] C. Anfinsen. Principles that govern the folding of proteins. *Science*, 181(96):223–230, 1973.

[2] S. Das and P. Suganthan. Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, 15(1):4–31, 2011.

[3] V. Feoktistov. *Differential Evolution: In Search of Solutions*. Springer, NY, 2006.

[4] A. Hagler and S. Lifson. Energy functions for peptides and proteins, II: The amide hydrogen bond and calculation of amide crystal properties. *Journal of the American Chemical Society*, 96:5319–5327, 1974.

[5] J. Handl, J. Knowles, R. Vernon, D. Baker, and S. Lovell. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. *Proteins*, 80(2):490–504, 2012.

[6] J. Lee, S. Wu, and Y. Zhang. Ab initio protein structure prediction. In *From Protein Structure to Function with Bioinformatics*, pages 3–25. Springer-London, 2009.

[7] B. Olson, K. De-Jong, and A. Shehu. Off-lattice protein structure prediction with homologous crossover. In *Proc. Conf. on Genetic and evolutionary computation - GECCO 2013*, pages 287–294, 2013.

[8] B. Olson and A. Shehu. Multi-objective stochastic search for sampling local minima in the protein energy surface. In *Proc. Inter. Conf. on Bioinformatics, Computational Biology and Biomedical Informatics - BCB 2013*, pages 430–439, 2013.

[9] Protein Data Bank. http://www.wwpdb.org.

[10] K. Price, R. Storn, and J. Lampinen. *Differential Evolution. A practical approach to global optimization.* Springer - Natural Comp. Series, 2005.

[11] C. Rohl, C. Strauss, K. Misura, and D. Baker. Protein structure prediction using Rosetta. *Methods in enzymology*, 383:66–93, 2004.

[12] Rosetta system. http://www.rosettacommons.org.

[13] J. Santos and M. Diéguez. Differential evolution for protein structure prediction using the HP model. *LNCS*, 6686:323–323, 2011.

[14] K. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, 1997.

[15] M. Sippl. Knowledge-based potentials for proteins. *Current Opinion in Structural Bio.*, 5:229 –235, 1995.

[16] R. Storn and K. Price. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.

[17] X. Zhao. Advances on protein folding simulations based on the lattice HP models with natural computing. *Applied Soft Comp.*, 8:1029–1040, 2008.