9

Some Modeling Issues for Protein Structure Prediction using Evolutionary Algorithms

Telma Woerle de Lima, Antonio Caliri, Fernando Luís Barroso da Silva, Renato Tinós, Gonzalo Travieso, Ivan Nunes da Silva, Paulo Sergio Lopes de Souza, Eduardo Marques, Alexandre Cláudio Botazzo Delbem, Vanderlei Bonatto, Rodrigo Faccioli, Christiane Regina Soares Brasil, Paulo Henrique Ribeiro Gabriel, Vinícius Tragante do Ó and Daniel Rodrigo Ferraz Bonetti
University of Sao Paulo
Brazil

1. Introduction

Many essential functions for life are performed by proteins and the study of their structures yields the ability to elucidate these functions in terms of a molecular view. (Creighton, 1992; Devlin, 1997) The interest in discovering a methodology for protein structure prediction (PSP) is of great interesti on many fields including drug design and carriers, disease mechanisms, and the food industry. In this context, several *in vitro* methods have been applied, as X-ray crystallography and nuclear magnetic resonance. Despite their relative success, both methods have their limitations. Conversely, the knowledge of the primary sequence of the amino acids of a protein can be achieved by a relatively simpler experimental measurement. From this information, one can in principle predict the three dimensional arrangement of its atoms, which has motivated the investigation of *ab initio* methods combining such initial knowledge with effective models (force fields) in order to predict the spatial structure of a protein (Bonneau & Baker, 2001; Hardin et al., 2002).

In fact, several computational methods for PSP are *semi ab initio* methodologies in the sense that they also use *prior* knowledge from both the sequence homology and the statistics found on protein databases [see e.g. (Miyazawa & Jernigan, 1985; Poole & Ranganathan, 2006)]. However, the use of these additional information restrict the search of protein structures that could be correctly predicted from the vast universe of proteins.

This chapter focuses on the development of a *pure ab initio* approach for PSP, not using prior information. In this context, evolutionary algorithms (EAs) have been investigated as a search method due to their flexibility to solve complex optimization problems. Our researches on EAs applied to PSP are twofold: 1) the investigation of more appropriate modeling of the physical and chemical interactions of a protein for the purpose of an optimization algorithm; 2) the development of computationally efficient EAs for PSP. Two important modeling issues have been poorly investigated in the literature related to the optimization techniques for PSP: a) the combined effects of the effective Hamiltonians based on force fields and the solvation free energy contribution (Section 3), and b) the use of

Source: Evolutionary Computation, Book edited by: Wellington Pinheiro dos Santos,
ISBN 978-953-307-008-7, pp. 572, October 2009, I-Tech, Vienna, Austria

multiple criteria to evaluate the predicted molecules since several physical interactions drive the folding process (Section 4). We show how both modeling issues can improve protein prediction quality.

We also present recently developed computational techniques to increase the efficiency of the algorithms for PSP. Algorithms using simple lattice models can work in a relatively small search space, however, they often generate a large number of unfeasible structures (with amino acid collisions). We present in this chapter lattice models that eliminate unfeasible solutions (Section 2). To increase the capacity of an EA to overcome local minimal of the potential energy, we propose an adaptation of the Self-Organizing Random Immigrants Genetic Algorithms for the PSP problem (Section 6). To work with large proteins, we explore computational strategies to enhance the efficiency of the calculi of the more complex energy functions (Section 5). Another strategy is the use of some heuristic about the proteins or its family to create the initial population of the EA, instead of the use of random solutions (Section 7).

Finally, this chapter shows how to combine the results from the set of above described researches in order to construct an *ab initio* PSP approach able to work with large proteins independently from prior information of similar structures (Section 8).

2. Advances using lattice models

Lattice models are simplified models of proteins which represent a conformation as a set of points in a grid. The simplest topologies of lattices are the squared lattice, for two dimensions, or the cubic lattice, for three dimensions. These models were originally employed in order to reduce the computational calculi (Dill, 1985; Unger & Moult, 1993). In this research field, they have been used to quickly evaluate the effect of parameter and operator of EAs, and, thus, motivated the development of new techniques in advanced models.

One of the most studied lattice models for protein folding is the *hydrophobic-hydrophilic model* (so-called HP model), where each amino-acid is classified in two classes: hydrophobic or non-polar (H), and hydrophilic or polar (P), according to their interaction with water molecules. (Chan & Dill, 1993a, 1993b) Moreover, each pair of amino acids in a conformation can be classified as *connected* or *neighbors*. Two amino acids from positions i and j in a sequence are connected if, and only if, $j = i + 1$ or $j = i - 1$. Notice that the number of amino acids is fixed. On the other hand, two amino acids in positions i and j are neighbors if the Euclidean distance between i and j is equal to 1. There are common features and assumptions behind such model with the classical Bragg-Willians and Flory-Huggins ones (Jönsson, B. et al, 1998).

The native state of a protein is a low-energy conformation. Thus, each pair of neighbors of H type contributes with a contact free energy -1. Then, the number of HH contacts is maximized in the native state. Despite the apparent simplicity of the model, finding the globally optimal conformation under HP model is an NP-Complete problem (Berger & Leighton, 1997), justifying the use of heuristic-based techniques for solving this problem. In the following, we present EAs developed for the PSP problem.

In the HP model, a protein conformation must be represented in a particular lattice; thus, each individual of the EA represents a conformation. In general, the fold is expressed as a sequence of *movements* into lattice. The position of the first amino acid is fixed and the other positions are specified by $n - 1$ movements for a sequence of n amino acids.

Two major schemes for representing the movements can be found in the literature:

- The *absolute representation* (AR) (Unger & Moult, 1993), where each new position is defined from the previous position. However, this representation allows movements of return, i.e., movements that annul the previous movement generating amino acid collision;
- The *relative representation* (RR) (Patton et al., 1995), where a movement is generated depending on the last movement in a way to avoid amino acid collision.

Since both representation do not avoid unfeasible solutions (with collisions), a penalty function assigns lower fitness values to these solutions during the evaluation stage. However, researches on EA representations for complex problems (Rothlauf, 2006) show that populations with feasible solutions have very slow convergence, since the unfeasible solutions dominated the evolutionary process with the increasing of the problem size. This phenomenon has been also verified in the PSP problem (Krasnogor et al., 1999; Cotta, 2003; Gabriel & Delbem, 2009).

In order to solve the problem of unfeasible solutions, Gabriel & Delbem (2009) propose a new approach using AR and a conformation matrix (C_M). This representation uses a matrix to decode AR of conformations. Each position of amino acid from a RR is indexed in a position of the matrix C_M representing the lattice. If there is already an amino acid in position (x,y,z) of C_M when decoding amino acid x, a collision is identified, x is replaced for an empty position of C_M and the corresponding movement in AR is properly updated.

To guarantee the efficiency of the decoding process, an array stores permutations of a set of relative movements (that are encoded using integers numbers) from a grid position. To repair the AR due to a collision, movements from the array are probed in C_M until finding a movement that avoids collision. If all possibilities in the array have been explored, the collisions are not repairable using local movements. Then, the individual (conformation) is eliminated, i.e., the building process is is canceled out and a new individual starts to be generated. For each collision a new array of permutations is constructed.

To analyzes the effect of C_M on the reduction of unfeasible solutions, we compare the quality of initial populations generated using C_M with random initial populations produced by classical EA based on AR. The usual fitness function for

Sequences	AR			AR + C_M		
	Best	Average	Worst	Best	Average	Worst
27 amino acids	0.50	-7.44	-3.40	3.20	0.57	0.00
64 amino acids	-7.80	-23.98	-62.80	6.90	2.00	0.00

Table 1. Comparison of the fitness value of the initial population using AR alone and AR with C_M.

HP models adds -1 for each collision in the conformation and +1 for each HH interaction (Gabriel & Delbem, 2009). Table 1 compares the average value the usual fitness of initial populations generated for 20 sequences (Unger & Moult, 1993; Patton et al., 1995): 10 with 27 amino acids and 10 with 64 amino acids. The percentage of feasible conformations in the initial population generated using AR. On the other hand, all conformations are feasible when AR and C_M are employed. In some populations, there are conformations very near to the global optimum.

3. Modeling the solvent effect for protein structure prediction

An aqueous solution is the environment where almost all important (bio)chemical reactions happen. It is central to several chemical, petrochemical, pharmaceutical and food engineering processes too (Eisenberg & Kauzmann, 1969; Ben-Naim, 1994; Devlin, 1997; Loehe & Donohue, 1997; Degrève & da Silva, 1999; Guillot, 2002; Dill et al, 2005; Levy & Onuchic, 2006). Proteins are just a type of solute found in this medium whose behavior is strongly affected by liquid water due to its intrinsic properties as a hydrogen-bonding solvent (Skaf & da Silva, 1994; Dill et al, 2005; Levy & Onuchic, 2006). In physiological conditions, the picture is far more complicated, principally because the presence of electrolytes (Na^+,K^+,Cl^-, etc.) (Jönsson, Lund & da Silva, 2007).

Water is well known to be a decisive factor on the protein conformational stability and determinative in processes such as the "protein folding" via the complex interplay especially of the solvent-protein and solvent-solvent interactions (Creighton, 1992; Eisenhaber & Argos, 1996; Devlin, 1997; Dill et al, 2005; Levy & Onuchic, 2006; Chaplin, 2008). In a classic view, these intermolecular interactions are given by electronic repulsion (whose origin is the Pauli exclusion principle), multipole-multipole interactions (often used as electrostatic interactions), induction (multipole-induced multipole interaction) and instantaneous induced multipole-induced multipole interactions (the London dispersion forces) (Israelachvili, 1991; Evans & Wennerström, 1994). The induction and dispersion interactions are known as the van der Walls interactions. In addition to these enthalpic contributions, due to the thermal effect, entropy has an important participation in these processes. Combined with the high affinity that water has for water, the final result is the so-called "hydrophobic effect", (Creighton, 1992; Garde et al., 1996; Jönsson et al, 1998; Levy & Onuchic, 2006; Chaplin, 2008) which is a key driven force for the protein folding.

In a molecular modeling approach, the initial challenge is to identify the main characteristics of the real physical system (in this case, a protein in solution), and to define how the intermolecular interactions and effects above mentioned may be replaced by suitable averages. This results in a mathematical expression that describes the potential energy of the system as a function of the separation distance between the species and is called an *effective* Hamiltonian model (EHM), (Friedman, 1977, 1981; da Silva, 1999) or, more commonly, the force field (FF) definition (Levitt, 1995; Ponder & Case, 2003; Mackerell Jr, 2004; Oostenbrink et al, 2004; Guvench & Mackerell Jr, 2008; Stone, 2008). We refer to this model as "effective", because it is not identical to the reality, but it is supposed to behave as the reality. Besides an atomist description, a vast diversity of coarse-grained models has also been reported (Tozzini, 2005; Monticelli et al, 2008).

Depending on the applications and questions to be answered, water may be modeled by different ways, from explicit (or molecular) to continuum (or implicit) solvent models (Friedman, 1981; Jorgensen et al, 1983; da Silva, Jönsson & Penfold, 2001; Guillot, 2002; Chen et al., 2008) (see Fig. 1). The main difference is the variables that *explicitly* enter the EHM and the computational costs. In the explicit models, the input variables are the coordinates and momenta of the solvent (i.e., a molecular model should be used), while the solvent only enters by an averaging over of its coordinates and momenta in the implicit models (i.e., water is replaced by its bulk static dielectric constant, ε_s). The latter model has a substantial reduction on cpu time with the price of loosing details that might be relevant for some cases. In this model level, the solvent in the immediate neighborhood of any solute is assumed to behave as the bulk solvent. It is an idealization widely used in Molecular Biology (Garcia-

Moreno, 1985; Bashford & Gerwert, 1992; da Silva et al., 2001; Chen et al., 2008) that captures *part* of the water properties, neglecting its molecular structure that is responsible principally for local, short-range, interactions and might be crucial in some cases (e.g. to study the hydration phenomenon).

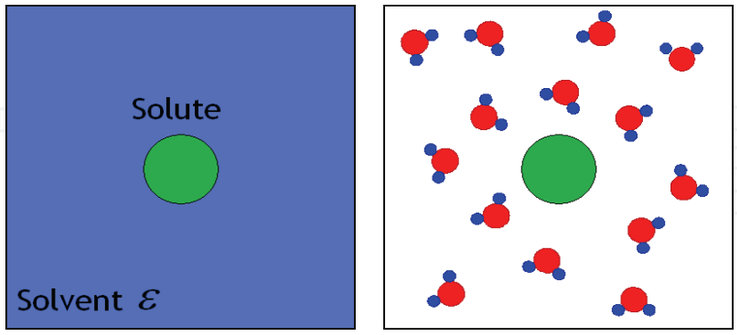


Fig. 1. A representation of a solute in two different water models: (a) A solute in a structure-less continuum model (implicit solvent model). Water only enters in the calculations by means of its bulk dielectric permittivity ε; (b) A solute molecule is surrounded by an explicit solvent model.

A large number of molecular models for water with different number of interaction sites based either on empirical or quantum mechanical methods have been developed during the last forty years, including rigid [e.g., SPC/E (Berendsen et al., 1987), TIP4P (Jorgensen et al, 1983), flexible [e.g., SPCFX (Teleman, O. et al., (1987)], dissociable [e.g., DCF (Halley et al., 1993)] and polarizable [e.g., SWM4-DP (Lamoureux et al., 2003), TIP4P/FQ (Rick, Stuart & Berne, 1994) models. Most water models assume site-site pair interactions in order to facilitate their application in numerical simulations. A common characteristic of them is to define a water molecule by a set of point sites. Fractions of electric charges are attributed to these sites and used to calculate the electrostatic pair potential energy (E_{ele}) according to Coulomb's law. The van der Waals forces are normally modeled by a particular case of Mie´s interaction potential known as Lennard-Jones (E_{LJ}) pair interaction potential (Verlet, 1967; Orea et al., 2008). The combination of these two terms (E_{ele}+E_{LJ}) is applied to the system constituents and gives the total interaction energy. Often a given model shows good behavior for a few properties and fails to describe others. Several reviews and comparisons between such models are available on the literature (van der Spoel et al., 1998; Wallqvist e Mountain, 1999; Guillot, 2002).

A compromise between the model details and the simulation costs is always a challenge task. To describe the hydrophobic effect, retaining the main characteristics of a real physical system (solute-solvent) and simulate it using reasonable cpu time, we need to find an adequate way to replace the exact calculation of the intermolecular interactions and effects by suitable averages, where the use of implicit solvent models is surely appealing. There are two basic models of implicit solvent normally used for this purpose: continuum electrostatic models and approaches based on solvent accessible surface (SAS) Variations of these models and combinations of them have also been proposed. (Garcia-Moreno, 1985; Bashford & Gerwert, 1992; Street & Mayo, 1998; da Silva et al., 2001; Snow et al, 2005, Chen et al., 2008).

Biomolecular systems are surrounded by solvent particles (water molecules, cations, and anions). In this environment, ions electrically interact with the molecule, which can have an

electric charge due to ionization mechanisms. (Jönsson et al., 2007) pH affects can be taken into account and give peculiar interactions. (da Silva et al, 2006; da Silva & Jönsson, 2009) The Poisson-Boltzmann equation (PBE) describes the electrostatic environment of a solute in solvent containing ions on a mean-field level of treatment. (Warwicker & Watson, 1982; da Silva, Jönsson & Penfold, 2001; Neves-Petersen & Petersen, 2003; Grochowski & Trylska, 2007; de Carvalho, Fenley e da Silva, 2008). It can be written as:

$$\nabla.\left[\varepsilon(r)\nabla\Phi(r)\right] = -4\pi\left[\rho(r) + qn_{+}\exp\left[\frac{-q\Phi(r)}{K_B T} - v(r)\right] - qn_{-}\exp\left[\frac{+q\Phi(r)}{K_B T} - v(r)\right]\right] \quad (6.1)$$

where the right term of the equation is the contribution of the fixed charges in the biomolecule, the other terms are the contribution of mobile positive and negative ions (treated here as point charges whose distribution around the central charged molecule obeys a Boltzmann distribution), $\varepsilon(r)$ is a position-dependent dielectric, $\Phi(r)$ is the electrostatic potential, $\rho(r)$ is the charge density of the solute, q is the charge of an ion, n_+ and n_- are densities of ion i at an infinite distance from the solute (bulk), K_B is the Boltzmann constant, $v(r)$ is 0 for accessible regions and infinite for unaccessible regions, and T is the temperature.
Analytical solutions of the PBE are possible only on ideal cases. One of these special situations is the infinite charged planar surface case. This example is given in details in refs. (Russel et al., 1989; Lyklema, 1991) and is usually described as the Gouy-Chapman case. (Usui, 1984) For biomolecular applications, numerical methods are necessary, and a number of different schemes have been proposed. (Lu et al, 2008) Nevertheless, despite the chosen numerical method, the PBE requires large memory resources and cpu time to be calculated without further approximations (e.g. to linearized it). (da Silva et al., 2001; de Carvalho et al., 2008) Although there are a large number of computational packages available (e.g. Melc, (Juffer, 1992) Delphi, (Sharp et al., 1998) APBS, (Holst et al., 2000) and MEAD (Bashford & Gerwert, 1992), the computational costs can still be a limitation and prohibitive when dealing with large molecules as proteins in applications that would require the systematic repetition of the same calculation for different macromolecular conformations. Critical investigations of the PBE are available elsewhere. (da Silva et al., 2001; de Carvalho et al., 2008)
The methods based on SAS arise experimentally by the linear relation between free energy and the surface area of a solute molecule (Hermann, 1972; Valvani et al., 1976; Amidon et al., 1975; Camilleri et al., 1988; Doucette e Andren, 1987; Dunn III et al., 1987; Snow et al, 2005). In this way, this method can directly provide the free energy of solvation. The continuum representation of solvent significantly reduces the cpu time. The SAS is the locus of points traced out by the inward facing part of the probe sphere, representing the solvent molecule that rotates over the van der Waals surface of the protein (see Fig. 2).
It is important to note that SAS solvent models also have limitations mainly related to:

- Viscosity: SAS and other implicit solvent models lack the viscosity, which affects the motion of solutes;
- Hydrogen-bonds with water: the directionality of the hydrogen bonds is missing in an implicit solvent. Moreover, bulk and buried water molecules are assumed to have the same behavior;
- Choice of model solvent: different atomic solvation parameters should be applied for modeling of the protein folding problem. These parameters should be derived from experimental coefficients involving organic molecules.

Nowadays, there are computational tools that calculate SAS and the free energy of solvation such as: GROMACS (Lindahl et al., 2001), POPS (Cavallo, 2003), Naccess (Hubbard e Thornton, 1993), STRIDE (Frishman e Argos, 1995), among others. In the following, we present preliminary results using SAS software based on the package STRIDE.

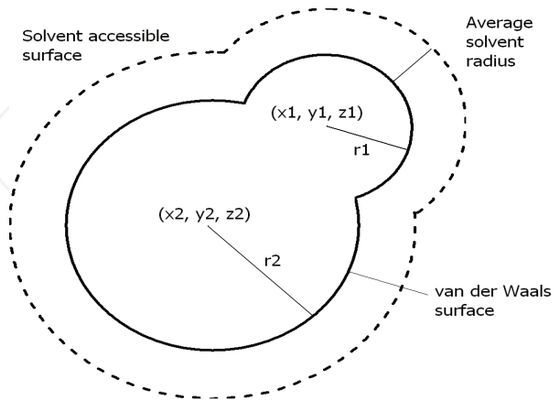


Fig. 2. A representation of van der Waals surface area and solvent accessible area surfaces (Richards, 1977).

The EA implementation for PSP, called ProtPred (Lima, 2007), was used to perform experiments to evaluate the effects of SAS and internal energies on the PSP. The potential energy functions are based on the CHARMm force fields of and solved by the Tinker package for molecular modeling (Ponder, 2001). The ProtPred uses the full-atom model (Cui et al, 1998; Cutello et al, 2005), representing the backbone and side-chain torsion angles (internal coordinates). The individuals (conformations) of the initial population are randomly generated with the angles in the constraint regions of each amino acid according to the Ramachandran map. The ProtPred code was run with 200 iterations and populations with 200 individuals were generated, using three recombination operators: i) two-point crossover, ii) uniform crossover, and iii) BLX-α (Deb et al., 2002). The algorithm uses three mutation operators: one changes all torsion angles of an amino acid by reselecting the values from the constraint regions; the others perform uniform mutation, but they differ from the use distinct step sizes.

The ProtPred minimizes a weighted objective function composed by energy functions. Fig. 3(a) shows the native protein structure of an acetylcholine receptor, obtained from PDB database (PDB id 1A11). Fig. 3(b) presents the structure obtained by ProtPred with the following weights for the objective function: 1.0 for wan der Waals, 0.5 for electrostatic, and zero for SAS. Fig. 3(c) shows the protein structure obtained changing the weights for SAS from zero to 0.001, indicating that SAS is relevant for PSP.

(a) (b) (c)

Fig. 3. Configurations of 1A11 protein.

Even though this protein is relatively small (25 aminoacids), the results encourage further works with this approach for modeling of the potential energy as a new criterion for evaluation in EAs to the PSP, considering not only the SAS, but also the internal energy.

4. Multiple criteria to evaluate predicted molecules

The PSP problem can be seen as a multi-objective problem, since it deals with several criteria involving energy functions estimating different aspects of intermolecular interactions (van der Waals, electrostatic, hydrogen-bond) and solvation free energies. A multi-criteria problem in general possesses the following characteristics:

- The determination of weights for an weighted function combining all the criteria is difficult;
- The criteria conflicts, i.e. the improvement of one objective in general involves the damage for other objective.

A protein structure with very low electrostatic energy may correspond to relatively high van der Waals energy. This kind of structure is in general inconsistent with the conformation of a real protein. Problems with conflicting objectives generally do not have a unique solution, but a solution set, called *Pareto*-optimal (Handl et al., 2006). This decade has produced relevant Multi-objective EAs, as the NSGA-II, which has been used to solve complex multi-objective problems. Unfortunately, even the NSGA-II losses performance for more than 3 objectives. For a large number of objectives, alternative solutions ought to be developed as the use of heuristics (Deb et al., 2006).

A first multiple criteria approach to PSP, called mo-ProtPred, was proposed in (Lima, 2006). This approach is based on NSGA-II (Deb, 2001), one of the main Multi-objective EA. The mo-ProtPred can work with three objectives without requiring weights for them.

The mo-ProtPred was applied to predict a transducin alpha-1 subunit (PDB id 1AQG). For this test, 500 generations with a population size equals to 400 were used. The objectives were functions corresponding to van de Waals, electrostatic, and hydrogen bond interaction energies. Fig. 4 illustrates polypeptides structures produced by the mo-ProtPred. Several of these structures are very similar to the native protein structure, displayed in Fig. 4(a). Table 2 shows the RMS values for each structure obtained.

Adequate structures for 1AQG were also achieved using a weighted function for the three objectives. Nevertheless, 27 different triples of weights were tested, 3 values (1.0, 0.5, 0.0) for each weight, and the ProtPred was executed for each triple. Thus, the ProtPred required much more running time than mo-ProtPred to achieve similar results.

	Conformations											
	a	b	c	d	e	f	g	h	i	j	k	l
RMS	0.00	3.83	4.09	3.83	3.77	4.09	4.06	3.93	3.77	4.01	4.14	3.83

Table 2. Results with 1AQG using mo-ProtPred.

5. Computation of van der Waals and electrostatic functions

The van der Waals potential energy models the induced and dispersion attraction among pairs of atoms and the electronic repulsion. It has as parameters the separation distance

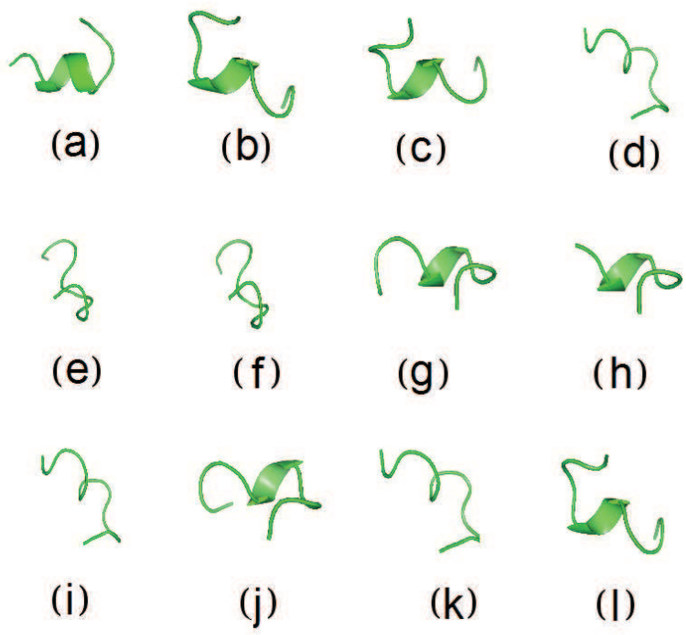


Fig. 4. Protein conformations of 1AQG obtained by mo-ProtPred.

between the atoms and van der Waals radii. At large distances, there is no attraction between pairs of atoms. However, when electron clouds of the atoms are close enough to overlap, there is a strong repulsion and its value exponentially grows. The energy is smaller when a balance point exists between the attractive and repulsive forces; this is known as van der Waals contact (Berg et al., 2002; Nelson, 2004). The Lennard-Jones 12-6 potential is frequently used to represent this interaction, because it models appropriately the attractive and repulsive forces and, from a numerical point of view, it can be efficiently implemented.

Together with the electrostatic interactions, van der Waals has considerable influence on the molecular structure. Studies have shown that the van der Waals forces contribute up to 65% of the total free energy of a protein (Becker, 2001).

In the evaluation of an individual to be used in an EA applied to PSP, there are several functions that contribute to the calculation of the minimum free energy of the protein. However, the computation of the van der Waals and electrostatic energies have time complexity $O(n^2)$, where n is the number of atoms. The interaction energy Eij is calculated for each atom pairs (i,j). The interactions in the molecule can be represented by a matrix E. Since $Eij = Eji$, E is an upper triangular matrix.

The computation time for both interaction energies corresponds to about 99% of the total EA execution time (Jain et al., 2009). It is therefore interesting to elaborate efficient algorithms and to use parallel processing wherever possible to reducing the total execution time.

There are classical methods for the parallelization of computations with upper triangular matrices. An obvious strategy is to distribute each row of the upper triangular matrix as a task for the processes. Therefore, using n processors, each one executing a task in parallel, the energy can be calculated in time $O(n)$, because the largest task would have n inter-atomic interactions to determine. On the other hand, some processors would have just some inter-

atomic interactions to compute, generating a non-uniform load balancing among processors. This can be reduced by combining the first row with the last row, the second row with the last but one row, and so on. This process produces $n/2$ tasks of size n (see Fig. 5).

The new individuals created by reproduction operators of EA have atoms with different coordinates from the parents. At each new generation, those coordinates must be sent to processors that calculate the electrostatics and van der Waals energies. The parallelization of these computations produces good results for large proteins, but not for small ones, as can be seen in Fig. 6. This figure shows the speedup achieved with the use of a given number of processors. The speedup is defined as the sequential execution time divided by the parallel execution time, and is equal to the number of processors in an ideal case (shown as a straight line in this figure). For a smaller protein, the computation time is on the order of the communication time for sending the coordinates to the processors, limiting the achieved speedup. For larger proteins, the computation time grows faster than the communication one, and the speedups are better. This is typical for parallel programs, which have overhead

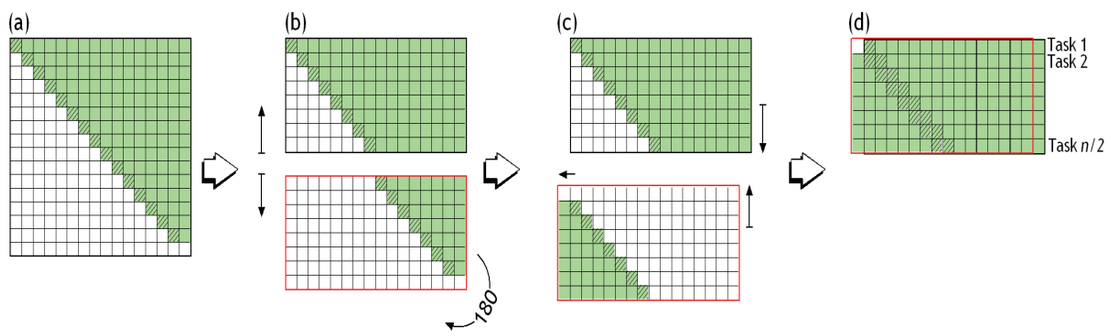


Fig. 5. (a) Upper triangular matrix of interaction. (b) Cut accomplished in the half of the matrix and the indication of the rotation of 180° of the second half. (c) Alignment of the second half to satisfy the dense matrix criterion. (d) New dense matrix of interactions with $n/2$ tasks of the same size.

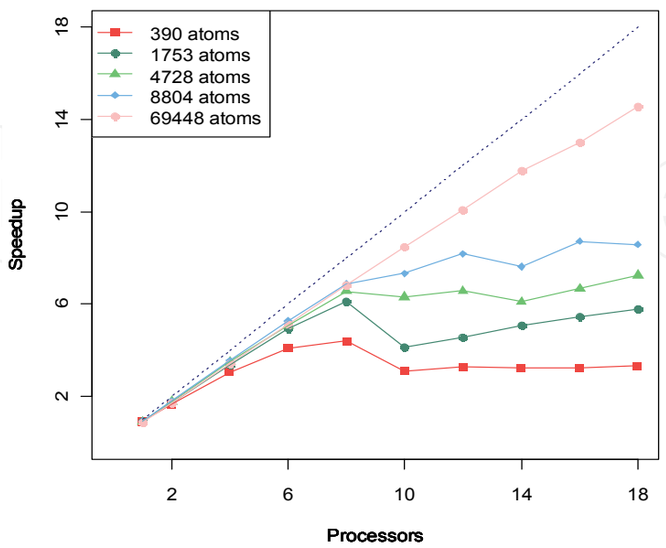


Fig. 6. The speedup reached by 5 different sizes of proteins to calculate van der Waals energy using 20 Athlon64-X2 processors.

costs that are independent of problem size (Quinn, 1994), and therefore work better for large instances. This phenomenon is also related to the Amdahl effect (Goodman & Hedetniemi, 1997). Since evaluations of different individuals are independent, rows of two or more dense matrices of interactions (Fig. 5) can be composed in larger tasks. This strategy can be especially adequate for the processing in GPUs (Graphic Processing Units) (Owens et al., 2008; Stone et al., 2007). The last generation of GPUs is capable of processing more than 3,500 matrices of 512x512 per second. This situation makes possible to compute energies of large molecules (hundreds of thousands of atoms) for several different configurations (individuals).

The computation time can also be improved using characteristics of the crossover operator used for the construction of new solutions. Since this operator preserves part of the information of each parent, the matrix of interactions of an offspring from crossover can be partly copied from that of its parents, as the energies depend only on the distances between the atoms, and those are preserved for pair of atoms that come from the same parent. Fig. 7 illustrates the areas of the upper triangular matrices that can be directly copied from the parents. The efficiency of this technique depends on the position of the crossover cut point. The worst case happens when the point is in the middle of the chromosome (see Fig. 8(a)).

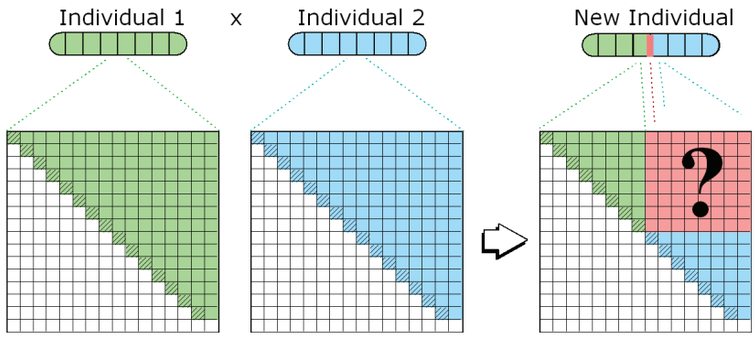
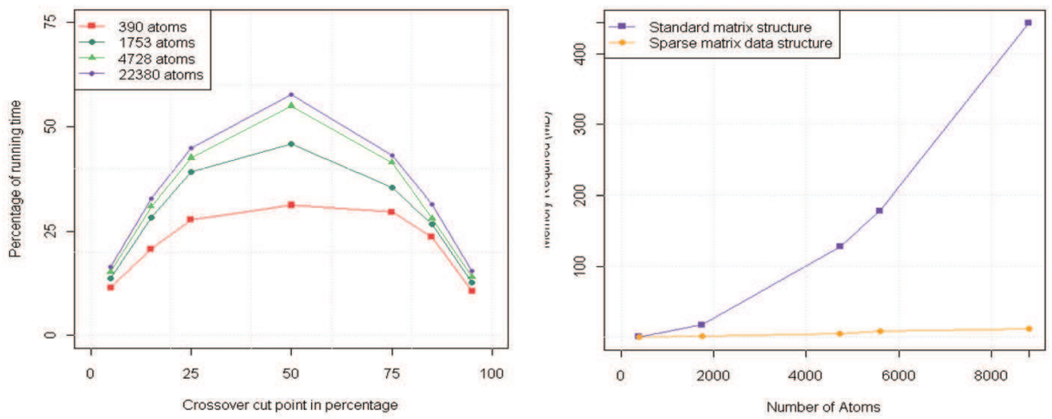


Fig. 7. Calculated values *Eij* of inter-atomic interactions for the parents are copied to the offspring. Only the pink area needs to be recalculated.



(a) (b)

Fig. 8. (a) Percentage of the required running time when *Eij*'s are copied from parent. (b) Required memory to save matrix *E* of an new individual.

Because both interaction types discussed here decay with the distance, to reduce the amount of computations it is usually defined a cutoff radius of 8Å for van der Waals and 13Å for electrostatic energies (Cui et al., 1998), and the computation of these energies for larger distances is avoided.

Taking this into account, we see that the matrix of interactions is in fact a sparse matrix. This helps to overcome a possible limitation for the computation of large proteins, as the interaction matrix would otherwise grow with $O(n^2)$. The data structure used is therefore important. Fig. 8(b) shows that the memory required for saving the whole matrix E increases quadratically with the size of the molecule, while using sparse matrix data structure the size increases more slowly.

Another optimization enabled by the use of a cutoff on the calculation of van der Waals and electrostatic energies makes possible to significantly reduce the amount of calculi for large proteins. One technique is the so called "Cell Lists" (Allen & Tildesley, 1987), which splits the space into cubic cells, as Fig. 9 illustrates (for the 2D case). Each cell has edge size equals to or larger than the cutoff radius. Thus, atoms in a cell interact only with other atoms in the same cell or in one of the neighbouring cells. This technique reduces the computation complexity from $O(n^2)$ to $O(n+m)$, where m is the number of cells with atoms, because a larger number of atoms in a protein is related with increased size, and therefore increased number of cells, instead of increased number of atoms per cell.

Fig. 10(a) shows the running cpu time required by the computation of the interactions without and with the cell-list approach. These result can also be improved by using an off-line procedure that previously compute the interaction energy between each possible pair of types of atoms (C, O, H, N, and S) for inter-atomic distances in a specified range (from a minimal distance to the maximal distance given by the cutoff radius) (Lodish et. at., 2003), see Fig. 10(b).

A performance enhancement for the computation of the potential energy has also been achieved using parallel solutions base on hardware like FPGAs (Field Programming Gate Arrays) (Wolf, 2004). The development of a specific solution in hardware is in general a hard task. One of the first researches using FPGAs for protein potential energy calculi did not reach a PC performance for the same problem. (Azizi et. al., 2004) In fact, this hardware solution was four time slower than the PC-base approach. Recent researches have achieved

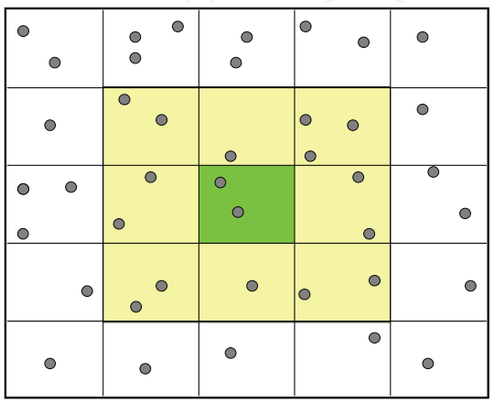
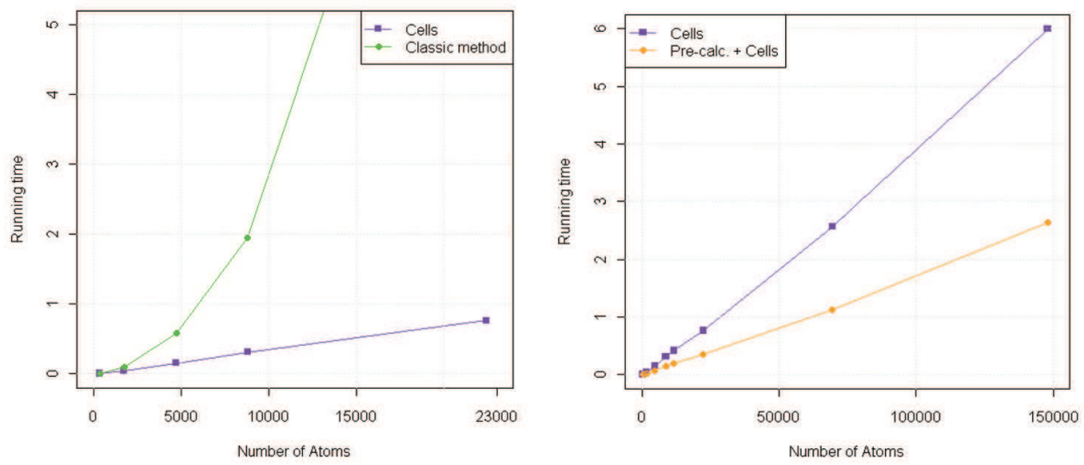


Fig. 9. 2D Cells, where the neighbour cells of the green cell are in yellow and the atoms are represented as grey balls.



(a) (b)

Fig. 10. Running cpu time necessary to calculate van der Waals potential: (a) Enhancement produced by the cell-list approach. (b) Improvement of performance combining cell-list and off-line calculi of *Eij*.

better performances by using faster memory bus and higher clock (Kindratenko & Pointer, 2006; Jain et. al., 2009). For example, Gu (2008) using a Virtex-5 (Xilinx, 2009) reached a speedup of 16.8.

The off-line calculi of inter-atomic energies (*Eij* 's) is another aspect that can be explored in FPGA solutions. The values *Eij* can be stored in lookup tables avoiding calculi with floating-point, which in general require large number of logic gates from FPGAs. Then, the available gates can be used to implement several lookup tables in parallel. The sum of the *Eij* 's can also be implemented in a parallel way. The implementation of lookup tables and the parallel sum are relatively simple, motivating more investigation of the use of FPGAs for PSP.

6. Simplified self-organizing random immigrants genetic algorithms

A common issue related to evolutionary computation and PSP is the premature convergence of the population towards local optimal solutions. The PSP problem has extremely complex search space complicating the determination of the global optima. In this case, a fast convergence is not a desired feature, at least on the former generations. In this sense, this section presents a strategy to insert population diversity in an organized way in Genetic Algorithms (GAs) with Random Immigrants (RIs), increasing the replacement rate whenever the population is becoming uniform.

The use of RIs introduces variation in the population [Grefenstette, 1992]. Thus, a GA with RIs can be more efficient in reaching better results than the conventional GA for the PSP problem [Tragante do O & Tinos, 2009]. However, the amount of RIs to be inserted at every generation has an important role in the GA evolution. The insertion of few RIs may not produce sufficient diversity in the population. On the other side, if a large number of individuals is replaced, the number of individuals of the population that explore the current best solutions may decrease, which can result in a smaller convergence rate.

In this way, a dynamic replacement rate, i.e. the algorithm decides the number of RIs for each population, can be useful to control the population diversity of the GA for the PSP problem. This is the objective of the Simplified Self-Organizing Random Immigrants (SSORIGA), which is based on the algorithm SORIGA presented in [Tinos & Yang, 2007] for dynamic optimization problems. In the original SORIGA, the worst individual and its neighbours are replaced by randomly-generated individuals and kept in a subpopulation in order to avoid the competition of RIs with other individuals to compose the new population, since the fitness of an individual from the subpopulation is probably worse than the ones from the remaining of the population.

Instead of creating a subpopulation, the SSORIGA inserts RIs automatically in the following generation as part of the normal population. Before generating the individuals for the next generation, the individual from the current population with the worst fitness is determined. If this individual is a RI just inserted, then the amount of RIs for the next generation is increased by 2. Otherwise, the number of RIs is restarted to 2. If the number of RIs reaches 70%, it is reset to 2. Fig. 11 presents the pseudocode of SSORIGA.

```
procedure generation ()
begin
  worst=find_worst_individual(population)
  if (lower_bound<index(worst)<upper_bound)
    totalrandom=totralrandom+2
              if (totalrandom>=0.7*population)
                totalrandom=2
              end_if
            else
              totalrandom=2
            end_if
            while (total_immigrants < totalrandom)
     son=new_individual()
     new_population.add(son)
               total_immigrants++
            end_while
  for(counter=percent;counter<pop_size;counter+2)
    father1 = tournament()      //selection of father 1
    father2 = tournament()      //selection of father 2
    son=father1.crossover(father2) //crossover
    son[0].mutation(mutation_rate)
    son[1].mutation(mutation_rate)
  end_for
 end.
```

Fig. 11. Pseudocode of SSORIGA.

The SSORIGA was adapted for the PSP as follows. All the Dihedral angles φ, ψ and χ_i (i the number of angles of the side chain) obtained from the protein structures found on the PDB database compose a set T of triples (φ, ψ, χ_i). Then, T is sorted according to φ generating a sequence S_T of triples. The representation of an individual (conformation of a protein of size n) is a sequence of n amino acids and pointers to (indices for) triples from T.

The mutation operator of SSORIGA only changes a pointer a little bit, automatically choosing a new triple in the neighbourhood of the pointer in S_T. The crossover operator is the usual with the one-point crossing-over. The selection operator is tournament selection, with 75% chance of the best individual being chosen to be one of the parents of the crossover.

The protein sequences *Crambin (PDB code 1CRN), Met-Enkephalin (PDB code 1PLW)* and *DNA-Ligand (PDB code 1ENH)* were used to evaluate the SSORIGA for the the PSP problem. The results indicated that SSORIGA was statistically better than the conventional GA approach and than RIGA with a fixed replacement rate of 6% or 10% depending on the protein. For other values of the replacement rate, SSORIGA was better than RIGA, indicating that SSORIGA was capable of finding the best replacement rate for RIs. It is important to note that:

- The mean number of replaced individuals in SSORIGA was between 3% and 5%;
- The small replacement rates occur in the first generations and the larger replacement rates take place periodically when the diversity is reduced.

In terms of potential energy values, the SSORIGA was able to reach much lower values than the standard GA. Comparing results with statistical tests, such as Student's T test, there is a probability of under 1.7% for all proteins of sampling error, considering as a final result the lowest energy value after the same number of evaluations of individuals for all algorithms.
In conclusion, the algorithm was suitable for reaching lower energy values than the standard GA. It can be applied to other domains, since the idea may be useful in other problems that require slight or heavy increase of diversity along time, such as Dynamic Optimization Problems.

7. Heuristics about protein to create the initial population

Although the number of sequenced proteins and their 3D structure determined experimentally grow systematically, the number of folds seems to be practically stabilized, since the year 2007, as shown by structural and topological classification of proteins databases, like SCOP and CATH. This scenario suggests that advances on computational methods, combined with the traditional techniques for theoretical prediction protein structure (Comparative modeling, Threading and *ab initio*), may improve substantially our capability for predicting protein structures. Algorithms that combine biological information from several databases with physical insights have proved to be a promising approach (Zhang, 2008). Each technique lonely has its specific strategies, advantages and limitations, as discussed in (Echenique, 2007) and illustrated in Table 3. Automated protein structure prediction is necessary because the protein folding process is not yet fully understood and experimental method (crystallographic and NMR) are relatively slow and financially expensive.
Echenique (2007) shows a schematic classification of methods for protein structure prediction, and discuss about experimental data and physical principles, emphasizing the need for more computer power and more accurate models. He concludes that, at the present stage, as more and more structural information is available, easier becomes the task for structural prediction. Indeed, in the last years we have seen increasing number and sophistication of biological databases: PDB, NCBI, Entrez, Dali Server (Holm et. al., 2008), SCOP, CATH, among others. These databases deal with experimental structural data reporting them directly (e.g. PDB), or presenting processed data information, such as in CATH.
The EA has been used for protein structure prediction as other sections discussed. The its population initialization is crucial for EA performance (Rahnamayan et al., 2007). Therefore, once considered that physical properties about protein folding, as proposed by Anfinsen (1973), in general lines, suggest that at physiological conditions the protein primary

Technique	Specific strategy
Comparative Modeling	Based on observation which proteins with similar sequences frequently share similar structure (Chotia & Lesk, 1986).
Threading	Try to identify similarities between 3D structure that aren't join by any significant sequence similarity. In other words, proteins often adopt similar folds despite no significant similarity sequence. Thus, can be to exist a limited number of protein folds in nature (Orengo et al., 2004).
Ab initio	Predict the protein native conformation from only its sequence aminoacids (Bourne & Weissig, 2003). No more information is needed.

Table 3. Technique of 3D structure prediction and their specific strategy.

sequence has all necessary information for its folding, Comparative Modeling (Table 3) may be used as a good method for generating the population initialization (Bourne & Weissig, 2003) because, beyond of being the easiest method for application, it is based in the following arguments:

- The protein structure is unique determined by its amino acid sequence (Anfinsen, 1973). Therefore, knowing the sequence it could, in principle, to obtain the corresponding native structure;
- Along the evolution process, proteins structures became resistant to changes. Therefore, similar sequences should correspond also to similar structures.

Furthermore considering the three prediction models, Comparative Modeling, has more information and accurate models than that its time computing required is lower (Echenique, 2007).

The next section will describe in more details Comparative Modeling. It is so clear that Comparative Modeling is good way for initial population for EA which will predict the 3D structure unknown. However, the difference between this initialization and the traditional approaches to Comparative Modeling is the latter choose only one and it is the solution. On the other hand, this section proposes that applying Comparative Modeling for to chose good individuals for the population instead of your random starting. The results intended are reducing the search space, number of individuals and generation.

Comparative Modeling

Greer (1981) argues that the central point to the traditional approach to Comparative Modeling is the insertion and deletion issue, which is generally treated by identifying *Structurally Conserved Regions* and doing local changes on the *Structurally Variable Regions*. This procedure may be split into eight steps (Orengo et al., 2004), namely,

- Identify one or more homologous sequences which have their structures known. These structures will be used as templates or parents;
- Align the target sequence to be modeled with the parents obtained in step 1;
- Determine the boundaries of the framework or Structurally Conserved Regions and the Structurally Variable Regions. Normally, Structurally Conserved Regions are loops;

- Inherit the *Structurally Conserved Regions* from the parents;
- Build the Structurally Variable Regions;
- Built the side-chains;
- Refine the model; and
- Evaluate errors in the model;

which are considered in more details as follows:

Step 1. The target sequence is searched through PDB database employing specific algorithms, such as Fasta and Blast to identify homologous[1] structures. Eventually, distant homologous can be identifying by PSI-Blast

Step 2. When the sequence identity is high (> 70%) the alignment is trivial. Moreover, when the identity is lower and the number of insertions and deletions is higher, it is very difficult to obtain a correct alignment which is fundamental for a good model (Orengo et al., 2004). Proteins of two (pair-wise) or more (multiple) sequence alignment may be directly compared, and this procedure is called Sequence Alignment. For pair-wise there are two types, global and local. The global alignment tries to align the entire sequences, using all their sequence characters. On other hands, local alignment tries to align sequence stretches and thus is created sub-alignments. Therefore, global alignments are used when sequences have quite similar and approximately the same length. Local alignments are employed when similar sequences have different lengths or share a conservation region or domain (Mount, 2004). Spaces may need to be inserted and there are called GAP (Pal et al., 2006). It is understood as multiple sequence alignment, that alignment of three or more sequences where each sequence columns represents the evolutionary changes in one sequence position (Mount, 2004). Simulated Annealing (Aart & Laarhoven, 1987) and Gibbs sampling (Lawrence et al., 1993), and particularly EA, have been used for multiple sequence alignment (Notredame & Higgins, 1996; Yokoyama et al. 2001; O'Sullivan et al., 2004). The difference about these EAs is their mutation operators (Pal et al., 2006)).

Step 3. *Structurally Conserved Regions* have all the same lengths are called core. The other regions, which differ structurally among parent sequences, are the *Structurally Variable Regions* (Orengo et al., 2004).

Step 4. This step can be divided into two situation:

- Single Parent – *Structurally Conserved Regions* are just copied from this parent and used as the model.
- Multiple Parents – distinct approaches may be used and choices depend on the modeler preferences, although the first step is always to fit structurally all multiple parents to one another.

Step 5. Normally, *Structurally Variable Regions* are loop regions. When the lengths of loops of the parent structures are different, they are built with lower accuracy when compared with the rest of the structure. Furthermore, even if the corresponding lengths of these regions are the same, they may adopt different structural conformations.

[1]*Homologous* refers to two or more sequences which have a common ancestor sequence in earlier evolutionary time. The ancestor is known after a complete alignment (Mount, 2004).

Step 6. There are various protocols to build the side chains: It can be simple like Maximum Overlap Protocol, in which side-chain torsion angles are inherited from their parent's side-chain, where possible, and additional atoms are built from a single conformation (Orengo et al., 2004). In the Minimum Perturbation Protocol Shih et. al. (1985) each substitution is guided by a rotation about the sidechain's torsion angles to relieve clashes. Another protocol, called Coupled Perturbation Protocol, developed by Mark Snow and Mario Amzel, is similar to the Minimum Perturbation Protocol, albeit the side chain torsion angles of structurally adjacent residues are also rotated.

Step 7. Model may be refined by Energy Minimization, a process in which all the atoms, governed by a previously specified force field, move until a conformation that represents a system minimum energy is reached. This issue is related to the Molecular Dynamics (MD) method (Frenkel & Smit, 1996). There are several popular software packages to run MD simulations, like Gromacs (Hess et al., 2008).

Step 8. A conventional measure of the modeling quality is done by applying the Root Mean Square Deviation (RMSD). Basically, RMSD shows how similar one structure is to another (Orengo et al., 2004). Usualy, the modeling quality may be tested first against a number of known protein structures.

Therefore, applying Comparative Model for starting the EA population may improve the efficiency of EA when compared by randomized starting population, since relevant initial information about unknown spatial structure of proteins are produced. Therefore, as argued by Zhang & Skolnick (2005), the PDB library is a systematically increasing contributor for the protein structure prediction problem. For example, such initial information can help EA get out of non-native local minimum energy; as well complementary information from specific biological databases may be used to build specific genetic operators.

8. Conclusions

It is noteworthy that the modelling, sampling and convergence properties might be a critical issue in the PSP. From a computation perspective based on EAS, different modeling issues for PSP were revised in this Chapter. Although lattice models are relatively simple, they are very appealing for EAS approaches where the computational efficiency can be highly improved, enabling the prediction of better protein structures. In fact, the data structure based on AR + C_M (Section 2) simplifies the objective function of lattice models since there is no need for an additional function penalizing amino acid collisions. As a consequence, the objective function uses only one criterion, i.e., the evaluation of the number of interactions between hydrophobic amino acids.

The hydrophobicity of protein is a measure of the interplay of the protein and solvent interactions. The objective function of the lattice models based on AR + C_M estimates this interaction. Thus, the EA using such model may also lead to a computationally efficient process in order to find protein conformations with more plausible solvent interaction.

The solvent effect on PSP is an important issue: for most cases, the solvation energy basically drives the process. Different alternatives on how to model the solvent have been pointed out on Section 3. Despite the fact that some protein structure were successfully obtained with models based on potential energy functions with no hydration free energy contributions, this is not a general rule. For a general protein case, the solvation free energy and interaction

potential energy functions are recommended for the proper prediction baring on mind the computational costs and efficiency. In order to use both contributions in PSP by EAs, two main issues should be adequately considered: 1) the computational efficiency of the energy calculi, and 2) an adequate manner to combine them. Section 5 presented strategies that reduce the running time to calculate van der Waals and electrostatic interaction energies from quadratic to linear. The off-line calculi of energies for a range of inter-atomic distances and the use of interaction energies previously calculated for parent solutions by the crossover operator can also enhance the computational efficiency.

The combined effect of different potential energies terms is dependent on the folding stage, i.e., the influence of a different term of the effective Hamiltonian may dominate the initial stage of the folding and another dominates the intermediate or the final stage. The influence of such terms on folding stages also depends on the protein molecule. In this context, the EAs enable us to simulate the effects of different combinations (weight set) of the energy functions involved in the PSP (Section 3). The most adequate weight set would reveal the key contribution of a Hamiltonian on the process contributing to enlighten the quantification of each physical mechanism behind the protein folding process.

Moreover, the mo-ProtPred can consider multiple criteria without previous knowledge of weights (Section 4) producing coherent protein structures. As an interesting consequence, the effects of each intermolecular interaction together with the solvation free energy can also be considered in PSP without previous information of the relative contribution of each Hamiltonian term on the folding.

The use of heuristics about proteins to create the initial population can improve significantly the performance EAs for PSP (Section 7). However, it may drive the algorithm for local optima, which may not correspond to an adequate protein structure. The SORIGA (Section 6) can be employed when heuristics population are used, reducing significantly the premature convergence for local optima. Thus, the initial-population heuristics combined with SORIGA should produce gains on convergence or even enable to work with larger proteins maintaining the quality of the prediction.

In short, the presented research attacked fundamental questions of the numerical sampling issues of the PSP. Moreover, it brought up solutions for each of these questions showing preliminary results and directions. In the literature, the questions clustered here have been the focus of several independent studies involving different areas of Science. The present chapter proposes some suggestions on how to combine the knowledge of diverse subjects to solve the PSP problem. There are issues that can be seen as details from on point of view but they are crucial from other perspective. Some of these aspects have been neglected in the development of a global solution for PSP on previous reviews.

In fact, hard problems can be characterized by the presence of several sets of highly interacting variables (Goldberg, 2002). One strategy to deal with such problems is to determine the interactions and adequately treat them. In PSP, the variables involve different research fields. Thus, a PSP modeling based on the integration of knowledge seems a promising path to achieve relevant advances.

9. References

Grefenstette, J. J. (1992). *Genetic algorithms for changing environments*, In: *Maenner, R., Manderick, B. (eds.) Parallel Problem Solving from Nature* 2, 137–144

Tragante do Ó, V. & Tinós, R. (2009). *Diversity Control in Genetic Algorithms for Protein Structure Prediction. VII Encontro Nacional de Inteligência Artificial (ENIA'2009)*, Bento Gonçalves, RS, Brazil

Tinós, R. & Yang, S. (2007). *A self-organizing random immigrants genetic algorithm for dynamic optimization problems. Genetic Programming and Evolvable Machines*, Vol. 8 No. 3, 255-286.

Lima, T.W. and Gabriel, P.H.R. and Delbem, A.C.B. and Faccioli, R.A. and da Silva, I.N. (2007). *Evolutionary algorithm to ab initio protein structure prediction with hydrophobic interactions*, IEEE CEC.

Cock, Peter J. A. and Antao, Tiago and Chang, Jeffrey T. and Chapman, Brad A. and Cox, Cymon J. and Dalke, Andrew and Friedberg, Iddo and Hamelryck, Thomas and Kauff, Frank and Wilczynski, Bartek and de Hoon, Michiel J. L., *Biopython*, Bioinformatics, 2009

Becker et al., 2001 Becker, O. M.; Jr., A. D. M.; Roux, B. & Watanabe, M. *Computational Biochemistry and Biophysics CRC,* 2001

Berg et al., 2002 Berg, J.; Tymoczko, J. & Stryer, L. *Biochemistry 5th ed W. H. Freeman,* 2002

Cui et al., 1998 Cui, Y.; Chen, R. S. & Wong, W. H. Protein Folding Simulation With Genetic Algorithm and Supersecondary Structure Constraints *Proteins: Structure, Function, And Genetics,* 1998, *31*, 247-257

Jain et al., 2009 Jain, A.; Gambhir, P.; Jindal, P.; Balakrishnan, M. & Paul, K. FPGA Accelerator for Protein Structure Prediction Algorithms *5th Southern Conference on Programmable Logic,* 2009, 123-128

Nelson, 2009 Nelson, D. L. & Cox, M. M. *Lehninger Principles Of Biochemistry Fourth Edition w. H. Freeman,* 2004

Owens et al., 2008 Owens, J. D.; Houston, M.; Luebke, D.; Green, S.; Stone, J. E. & Phillips, J. C. GPU Computing *Proceedings of the IEEE,* 2008, *96*, 879-899

Stone et al., 2007 Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G. & Schulten, K. *Accelerating Molecular Modeling Applications with Graphics* Processors Journal of Computational Chemistry, 2007, *28*, 2618-2640

Cotta, C. (2003). *Protein structure prediction using evolutionary algorithms hybridized with backtraking*, International Work-conference on Artificial and Natural Neural Networks, pp. 321-328, 2003, Lecture Notes in Computer Science

Berger, B. & Leigthon, T. (1997). *Protein folding in the hydrophobic-hydrophilic model (HP) is NP-complete*, Journal of Computational Biology, 5, 1, 27-40

Dill, K.A. (1985). *Theory for the folding and stability of globular proteins. Biochemistry*, 24, 6, 1501-1509

Gabriel, P.H.R. & Delbem, A.C.B (2009). *Representations for Evolutionary Algorithms applied to Protein Structure Prediction Problem using HP Model*, Brazilian Symposium on Bioinformatics (In Press)

Krasnogor, N.; Hart, W.E.; Smith, J. & Pelta, D.A. (1999). *Protein Structure Prediction with Evolutionary Algorithms, Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1596-1601, Orlando, Florida, 1999, Morgan Kaufmann*

Patton, A.L.; Punch III, W.F. & Goodman, E.D. (1995). *A standard GA approach to native protein conformation prediction*, Proceedings of the Genetic and Evolutionary Computation Conference, pp. 574-581, San Francisco, CA, 1995, Morgan Kaufmann

Rothlauf, F. (2006). *Representation for Genetic and Evolutionary Algorithms*, Springer-Verlag

Unger, R. & Moult, J. (1993). *A genetic algorithm for 3D protein structure simulations*. Proceedings of Fifth Annual International Conference on Genetic Algorithm, San Francisco, CA, 1993

Bashford, D. & Gerwert, K. (1992). *Electrostatic calculations of the pKa values of ionizable groups in bacteriorhodopsin*, J. Mol. Biol., 224, 473-486.

Ben-Naim, A. (1994). *Solvation: from small to macro molecules*, Curr. Opin. Struct. Biol., 4, 264-268.

Berendsen, H. J. C. , Grigera, J. R. & Straatsma, T. P. (1987). *The Missing Term in Effective Pair Potentials*. J. Phys. Chem., 91, 6269-6271.

Bonneau, R. & Baker, D. (2001). Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.*, 31, 173-189.

Chaplin, M. F. (2008). *Roles of Water in Biological Recognition Processes, Wiley Encyclopedia of Chemical Biology*, 2008, 1-8.

Chan, H. S. & Dill, K. A. (1993*a). The protein folding problem*. Physics Today, 46, 24-32.

Chan, H. S. & Dill, K. A. (1993b). *Energy landscapes and the collapse dynamics of homopolymers,* J. Chem. Phys., 99, 2116-2127.

Chen, J., Brooks III, C.L. & Khandogin, J. (2008). *Recent advances in implicit solvent-based methods for biomolecular simulations*, Current Opinion in Structural Biology, 18, 140-148.

Creighton, 1992. Creighton, T. E. *Protein Folding,* W. E. Freeman and Company, New York, 1992.

Da Silva, F.L.B., 1999. Da Silva, F.L.B. *Statistical Mechanical Studies of Aqueous solutions and Biomolecular Systems*, Lund University, SLU, Alnarp, 1999.

Da Silva, F.L.B., Jönsson, B. & Penfold, R. (2001). *A critical investigation of the Tanford-Kirkwood scheme by means of Monte Carlo simulations, Prot. Science*., 10, 1415-1425.

Da Silva, F.L.B., Lund, M, Jönsson, B. & Åkesson, T. (2006). *On the Complexation of Proteins and Polyelectrolytes*, *J. Chem. Phys. B*. 110, 4459-4464.

Da Silva, F.L.B., & Jönsson, B. (2009). *Polyelectrolyte–protein complexation driven by charge regulation, Soft Matter*, no prelo.

Degrève, L. & da Silva, F.L.B. (1999). *Structure of concentrated aqueous NaCl solution: A Monte Carlo study, J. Chem. Phys.,* 110, 3070-3078.

Devlin, 1997. Devlin, T. M. *Textbook of Biochemistry with Clinical Correlations,* Wiley-Liss, New York, 1997.

Dill, K. A., Truskett, T. M., Vlachy, V. & Hribar-Lee, B. (2005). *Modeling Water, the hydrophobic effect, and ion solvation*, Annu. Rev. Biophys. Biomol. Struct., 34, 173-179.

Eisenberg & Kauzmann, 1969. Eisenberg, D. & Kauzmann, W. *The Structure and Properties of Water,* Oxford University Press, New York, 1969.

Eisenhaber, F. & Argos, P. (1996). *Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation*, Protein Eng., 9, 1121-1133.

Evans & Wennerström, 1994. Evans, D.F. & Wennerström, H. *The Colloidal Domain,* VCH Publishers, New York, 1994.

Friedman, H. L. (1977). *Introduction*, Faraday Discuss. of the Chem. Soc., 64, 7-15.

Friedman, H. L. (1981). *Electrolyte solutions at equilibrium, Ann. Rev. Phys. Chem*., 32, 179-204.

Garcia-Moreno, B. (1985). *Probing Structural and Physical Basis of Protein Energetics Linked to Protons and Salt*, Methods in Enzymology, vol. 259, 512-538.

Garde, S. et al. (1996). Garde, S., Hummer, G. , Paulaitis, M. E., Garcia, A. E. & Pratt, L. R.. (1996). *Origin of entropy convergence for hydrophobic hydration and protein folding*, Phys.Rev. Letts., 77, 4966-.

Grochowski, P & Trylska, J. (2007). *Continuum Molecular Electrostatics, Salt Effects, and Counterion Binding – A Review of the Poisson–Boltzmann Theory and Its Modifications*, Biopolymers, 89, 93-113.

Guillot, B. (2002). *A reappraisal of what we have learned during three decades of computer simulations on water*, Journal of Molecular Liquids, 110/1-3, 219-260.

Guvench, O. & Mackerell Jr, A. D., 2008. *Comparison of Protein Force Fields for Molecular Dynamics*

Simulations, In: *Andreas Kukol (ed.) Methods in Molecular Biology: Molecular Modeling of Proteins*, vol. 443, 63-88, Humana Press, Totowa, NJ, 2008.

Hardin, C., Pogorelov, T.V., Luthey-Schulten, Z. (2002). *Ab initio protein structure prediction*, Current opinion in structural biology, 12, 176-181.

Halley, J. W., Rustad, J. R. & Rahman., A. (1993). *A polarizable, dissociating molecular dynamics model for liquid water, J. Chem. Phys*., 98, 4110-4119.

Holst, M., Baker, N., Wang, F. (2000) *Adaptive multilevel finite element solution of the Poisson-Boltzmann equation I. Algorithms and examples*. J. Comput. Chem. 21, 1319-1342.

Israelachvili, J., 1991. Israelachvili, J. *Intermolecular and Surface Forces, 2nd. Ed.,* Academic Press, London, 1991.

Jönsson, B. et al, 1998. Jönsson, B., Lindman, B., Holmberg, K. & Kronberg, B. *Surfactants and Polymers in Aqueous Solution,* John Wiley, Chichester, 1998.

Jönsson, B., Lund, M. & da Silva, F.L.B., 2007. Electrostatics in Macromolecular Solution, In*: Eric Dickinson and Martin E. Leser (eds.) Food Colloids: Self-Assembly and Material Science.*, Chapter 9, 129-154, Royal Society of Chemistry, 2007.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). *Comparison of simple potential functions for simulating liquid water*, J. Chem. Phys., 79, 926-935.

Juffer, A. H (1992). Melc – *The Macromolecular Electrostatics Computer Program; Laboratory of Physical Chemistry*, University of Groningen: Groningen, The Netherlands, 1992.

Lamoureux, G., Mackerell Jr, A. D., & Roux, B. (2003). *A simple polarizable model of water based on classical Drude oscillators*, J. Chem. Phys., 119, 5185-5197.

Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). *Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution*, Computer Physics Communications, 91, 215-231.

Levy, Y. & Onuchic, J. N. (2006). *Water Mediation in Protein Folding and Molecular Recognition,* Annu. Rev. Biophys. Biomol. Struct., 35, 389-415.

Loehe, J. R. & Donohue, M. D. (1997). *Recent advances in modeling thermodynamic properties of aqueous strong electrolyte system*, AIChE J., 43, 180-195.

Lu, B. Z., Zhou, Y. C., Holst, M. J. & J.A. McCammon. (2008). *Recent Progress in Numerical Methods for the Poisson-Boltzmann Equation in Biophysical Applications, Commun. Comput. Phys.*, 3, 973-1009.

Lyklema, J (1991). *Fundamentals of Interface and Colloid Science*. Academic Press, San Diego.

Mackerell Jr, A. D. (2004). *Empirical Force Fields for Biological Macromolecules: Overview and Issues*, J. Comput. Chem., 25, 1584-1604.

Miyazawa, S. & Jernigan, R.L. (1985). *Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation,* Macromolecules, 18, 534-552.

Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R. G. , Tieleman, D. P. & Marrink, S-J. (2008). *The MARTINI Coarse-Grained Force Field: Extension to Proteins*, J. Chem. Theory and Comput., 4, 819-834.

Neves-Petersen, M.T. & Petersen, S.B. (2003). *Protein electrostatics: A review of the equations and methods used to model electrostatic equations in biomolecules – Applications in biotechnology*, Biotechnology Annual Review, 9, 315-395.

Oostenbrink, C., Villa, A, Mark, A.E. & van Gunsteren, W.F. (2004). *A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6, J. Comput. Chem.*, 25, 1656-1674.

Orea, P., Reyes-Mercado, Y. & Duda, Y. (2008). *Some universal trends of the Mie(n,m) fluid thermodynamics, Phys. Letts*. A, 372, 7024–7027.

Ponder, J. W. & Case, D.A. (2003). *Force Fields for protein simulations, Adv. Prot. Chem.*, 66, 27-85.

Poole, A. M. & Ranganathan, R. (2006). *Knowledge-based potentials in protein design*, Current Opinion in Structural Biology In Membranes / Engineering and design, 16, 508-513.

Rick, S. W., Stuart, S. & Berne, B. J. (1994). *Dynamical Fluctuating Charge Force Fields: Application to Liquid Water*, J. Chem. Phys., 101, 6141−6156.

Russel, W. B. , Saville, D. A. & Schowalter, W. R. (1989). *Colloidal Dispersions*, Cambridge University Press, Cambridge.

Sharp, K. A., Nicholls, A., Sridharan, S. (1998). Delphi – *A Macromolecular Electrostatics Modeling Package*; Columbia University: New York, 1998.

Snow, C. D. D., Sorin, E. J. J., Rhee, Y. M. M. , Vijay & Pande, S. S. (2005). *How Well Can Simulation Predict Protein Folding Kinetics and Thermodynamics?*, Annu. Rev. Biophys. Biomol. Struct., 34, 43-69.

Skaf, M. S. & da Silva, F.L.B. (1994). *Explorando as propriedades moleculares de solventes*, Química Nova, 17, 507-512.

Stone, A.J. (2008). *Intermolecular Potentials,* Science, 321, 787-789.

Street, A.G. & Mayo, S. L. (1998). *Pairwise calculation of protein solvent-accessible surface areas*, Folding & Design, 3, 253-258.

Teleman, O., Jönsson, B. & Engström, S. (1987). *A molecular dynamics simulation of a water model with intramolecular degrees of freedom*, Mol. Physics, 60, 193-203.

Tozzini, V. (2005). *Coarse-grained models for proteins*, Current Opinion in Structural Biology, 15, 144-150.

Usui, S. (1984). *Electrical double layer*. In A. Kitahara and A. Watanabe, editors, *Electrical Phenomena at Interfaces : Fundamentals, Measurements, and Applications*, 15-46, New York, 1984. Marcel Dekker, Inc.

van der Spoel, D., van Maaren, P. J. & Berendsen, H. J. C. (1998). *A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field*, J. Chem. Phys., 108, 10220-10230.

Warwicker, J. & Watson, H. C. (1982). *Calculation of the electric potential in the active site cleft due to α-helix dipoles*, J. Mol. Bio., 157, 671-679.

Richards, F. M. (1977). *Areas, volumes, packing and protein structure*. Annu Rev Biophys Bioeng., 6:151–176.

Aart, E. & Laarhoven, V. P. (1987). *Simulated Annealing: A Review of Theory and Applications*, Norwell, MA Kluwer, 9789027725134

Anfinsen, C. B., (1973). *Principles that Govern the Folding of Protein Chains*, Science, 181, 4096

Bourne, P. E. & Weissig, H., (2003) *Structural Bioinformatics*, Wiley-Liss, Inc, 9780471202004

Chothia, C. & Lesk, A. M., (1986). *The relation between the divergence of sequence and structure in proteins*, EMBO J. 5, 823–826

Echenique, P. (2007). *Introduction to protein folding for physicists*, Contemporary Physics, 48, 2, 81-108

Frenkel, D. & Smit, B., (1996). *Understanding Molecular Simulation*, Academic Press, INC, 9780122673511

Greer, J. (1981).*Comparative model-building of the mammalian serine proteases*, Journal of Molecular Biology, 153, 4, 1027-1042

Hess, B., Carsten, K., van der Spoel, D. & Lindahl, E. (2008). *GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation*, J. Chem. Theory Comput,. 4, 3, 435–447

Holm, L., Kääriäinen, S., Rosenström, P. & Schenkel, A. (2008) . *Searching protein structure databases with DaliLite v.3*, Bioinformatics Applications Note, 24, 23, 2780–2781

Joseph-McCarthy, D., Petsko, A.G. & Karplus, M. (1995). *Use of a minimum perturbation approach to predict TIM mutant structures*, Protein Eng. 8, 11, 1103-1115

Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A. & Wootton, J. (1993) *Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment*, Science, 262, 208–214

Mount, D. W. (2004). *Bionformatics Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Pr, 9780879697129

Moult, J., Fidelis, K., Rost, B., Hubbard, T. & Tramontano, A.(2005) *Critical Assessment of Methods of Protein Structure Prediction (CASP) – Round 6*, PROTEINS: Struct. Funct. Bioinf., 7, 3-7

Notredame, C. & Higgins, D. G., (1996). *SAGA: Sequence alignment by genetic algorithm*, Nucleic Acids Res., 24, 8, 1515–1524

Orengo, C. A., Jones, D. T. & Thornton, J. M., (2004). *Bioinformatics genes, Proteins & Computers*, Bios Scientic Publishers Limited, 9781859960547

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G.., & Notredame, C. (2004) *3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments*, Journal of Molecular Biology, 340 2, 385-395

Pal S.K., Bandyopadhyay, S. & Ray, S.S. (2006). *Evolutionary computation in bioinformatics: a review*, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 36, 5, 601-615

Rahnamayan S.,Tizhoosh, H. R. & Salama, M. M.A. (2007). *A novel population initialization method for accelerating evolutionary algorithms*, Computers and Mathematics with Applications, 53, 1605–1614

Shih, H. H. L., Brady, J., & Karplus, M. (1985). *Structure of proteins with single-site mutations: A minimum perturbation approach,* Proc. Nati. Acad. Sci., 82, 1697-1700

Yokoyama, T., Watanabe T., Taneda A. & Shimizu T., (2001). *A Web Server for Multiple Sequence Alignment Using Genetic Algorithm*, Genome Informatics, 12, 382–383

Zhang, C. & Wong, A. K. C., (1997). *A genetic algorithm for multiple molecular sequence alignment*, Bioinformatics, 13, 565–581

Zhang, Y. & Skolnick J. (2005). *The protein structure prediction problem could be solved using the current PDB library*, Proc Natl Acad Sci, 102, 1029-1034

Zhang, Y. (2008). *Progress and challenges in protein structure prediction*, Current Opinion in Structural Biology, 18, 342–348

Goldberg, D. E. (2002), The Design of Innovation, Lessons from and for Competent Genetic Algorithms, Kluwer Academic Publishers, Norwell, MA, USA

Deb, K. (2001), Multi-Objective Optimization using Evolutionary Algorithms, Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Chichester.

Dill, K.A. (1995). Principles of Protein folding – A perspective form simple exact models, Protein Science, 4, :pp. 561-602

Dill, K. A. (2005). Truskett, T. M., Vlachy, V., Hribar-Lee, B. Modeling Water, the hydrophobic effect, and ion solvation, Annu. Rev. Biophys. Biomol. Struct., 34, 173-179

Verlet, L. (1967). Computer experiments on classical fluids.i. thermodynamical properties of lennard-jones molecules. Phys. Rev., v. 159, p. 98

Valvani, S. C., Yalkowsky, S. H.; Amidon , G. L. (1976). Solubility of nonelectrolytes in polar solvents: V. estimation of the solubility of aliphatic monofunctional compounds in water using the molecular surface area approach. J. Phys. Chem., p. 829

Wallqvist, A., Mountain , R. D. (1999). Molecular models of water: Derivation and description. Reviews in Computational Chemistry, v. 13, p. 183–247

Hermann, R. B. (1972). Theory of hydrophobic bonding ii. the correlation of hydrocarbon solubility in water with solvent cavity surface area. J. Phys. Cm., v. 76, p. 2754

Doucette, W. J., Andren, A. W. (1987). Correlation of octanol/water partition coefficients and total molecular surface area for highly hydrophobic aromatic compounds. Environ. Sci. Technol., v. 21, p. 821

Dunn III, W. J., Koehler, M. G., Grigoras , S. (1987). The role of solvent-accessible surface area in determining partition coefficients. J. Med. Ckem., v. 30, p. 1121

Camilleri , P., Watts , S. A., Boroaston, J. A. (1988). A surface area approach to determination ofpartition coefficients. I. C/rem. Sue. Perkin Trans., v. 11, p. 1699.

Lindahl , E., Hess , B., Spoel, D. (2001). Gromacs 3.0: a package for molecular simulation and trajectory analysis. J. Mol. Mod., p. 306–317

Hubbard , S.; Thornton , J. (1993). Naccess: computer program. Department of Biochemistry and Molecular Biology, University College London

Frishman, D., Argos, P. (1995). Knowledge-based secondary structure assignment. Proteins: structure, function and genetics, v. 23, p. 566–579

Evolutionary Computation

Edited by Wellington Pinheiro dos Santos

ISBN 978-953-307-008-7
Hard cover, 572 pages
Publisher InTech
Published online 01, October, 2009
Published in print edition October, 2009

This book presents several recent advances on Evolutionary Computation, specially evolution-based optimization methods and hybrid algorithms for several applications, from optimization and learning to pattern recognition and bioinformatics. This book also presents new algorithms based on several analogies and metafores, where one of them is based on philosophy, specifically on the philosophy of praxis and dialectics. In this book it is also presented interesting applications on bioinformatics, specially the use of particle swarms to discover gene expression patterns in DNA microarrays. Therefore, this book features representative work on the field of evolutionary computation and applied sciences. The intended audience is graduate, undergraduate, researchers, and anyone who wishes to become familiar with the latest research work on this field.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Telma Woerle de Lima, Antonio Caliri, Fernando Luis Barroso da Silva, Renato Tinos, Gonzalo Travieso, Ivan Nunes da Silva, Paulo Sergio Lopes de Souza, Eduardo Marques, Alexandre Claudio Botazzo Delbem, Vanderlei Bonatto, Rodrigo Faccioli, Christiane Regina Soares Brasil, Paulo Henrique Ribeiro Gabriel, Vinicius Tragante do O and Daniel Rodrigo Ferraz Bonetti (2009). Some Modeling Issues for Protein Structure Prediction Using Evolutionary Algorithms, Evolutionary Computation, Wellington Pinheiro dos Santos (Ed.), ISBN: 978-953-307-008-7, InTech, Available from: http://www.intechopen.com/books/evolutionary-computation/some-modeling-issues-for-protein-structure-prediction-using-evolutionary-algorithms

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen