Protein folding with local propensity using GA

Carlos Leandro
[miguel.melro.leandro@gmail.com]
Departamento de Matemática,
Instituto Superior de Engenharia de Lisboa, Portugal.

April 11, 2016

Abstract

This work presents my results about the usefulness of rotamer libraries to predict protein folding. The most important feature of this strategy is the possibility of modelling protein folding without explicitly treating every atom in the problem. Here I try to evaluate the importance or relevance of two empirical energy function to predict the folding. These functions are defined using potentials extracted from Ramachandran plots (Ramakrishann 1965), studies of this plots (McCammon 1984) show that they reflect the local interactions of free energy. The protein conformation is determined from a balance between local interactions (in each amino acid) and non-local ones encoded in the protein potential energy function as statistical potentials. I explored two approximations to the free energy landscape of several proteins using the Genetic Algorithm optimization solver available in the Matlab Optimization Toolbox, trying predict a stable conformation based on the proteins first base. For that several proteins (2FKL, 1PEN, 1NOT, 1FXD) were selected from the PDB, extracting its first base, the Ramachandran plot for each of its residuum were computed and used them to approximate the protein free energy landscape. Here this empirical energy is used as a fitness function in a Genetic Algorithm Optimization procedure to predict the dihedral angles for a minimum energy conformation. For evaluate the quality of this prediction, the predict dihedral angles are compared to the protein native values. I used for similarity measurement the mean square error (MSE) between the vector of predicted dihedral angles and the value of this angles in a stable conformation.

Contents

1 Introduction

Biological proteins are polymeric chains build from amino acid monomers. These amino acids contain five chemical components: a central α-carbon (C_α), an α-proton (H), an amino functional group ($-NH$), a carboxylic acid functional group ($-COOH$), a side chain group (R). These amino acids combine to become proteins through an energy-driven combination. This result in the creation of a peptide bond between the two amino acids, and repeating the process creates a polypeptide containing several peptide bonds. These peptide bonds behave like a partial double bond, which have restricted rotation about the bond. This restriction results in a stable peptide plane. These peptide planes are repeating units that exhibit constant structures in the protein and reduce the number of degrees of freedom in the protein. The polypeptide chain is intrinsically flexible because many of the covalent bonds that occur in its backbone and sidechains are rotationally permissible.

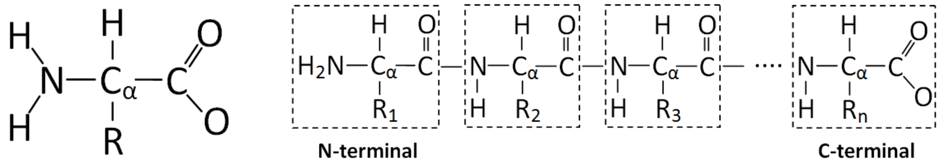


Figure 1: Fig. 1 Amino acids are molecules containing an amine group (H_2N-), a carboxylic acid group ($-COOH$), and a side-chain (R) that is pecific to each amino acide. The first carbon that attaches to a functional group is named apha-carbon (C_α). Fig. 2 Every peptide has a N-terminus residue and a C-terminus residue on the ends of the peptide.

Geometric relationship involving atoms in the polypeptide fully define the thee-dimensional proteins structure. The relationships consist of bond lengths, bond angles, dihedral angles and improper dihedral angles. The primary contributions from these parameters, which determine overall polypeptide structure, are the dihedral angles. Typically, the peptide plane remains relatively rigid during protein dynamics such that the bond lengths and bond angles remain constant, due the large energy cost for its deformation. As a result, the dihedral angles are the essential degrees of freedom that dictate the position of the polypeptide backbone atoms, defining the protein secondary structure. In addition polypeptide-solvent interactions and a number of important interactions between non-bonded atoms of the polypeptide help to determine the native structure of a protein.

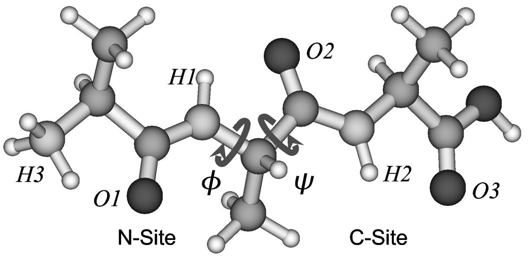


Figure 2: Backbone dihedral angles in the molecular structure of trialanine (Altis 2008).

Results indicated that the torsional motion is predominantly local in character.

Based in this facts, in this work, the protein model has been simplified as a constrained multibody system, and the overall dynamic is described by their backbone dihedral angles. The most important feature of the formulation proposed is the possibility of modelling protein folding without explicitly treating every atom in the problem. Using this quasicontinuum approach, must degrees of freedom are eliminate, and energy calculations are expedited. For that we used the refinement proposed in Fig. 3, for a generic protein defined by n amino acid, using a constrained multibody system with $2n+1$ bodies $b_1, b_2, \ldots, b_{2n+1}$, linked together using revolute joints, with dihedral angles $\beta_1, \beta_2, \ldots, \beta_{2n}$. This ensures a full atomic detail in regions of the protein where it is required. Since this polypeptide chain consists of a large number of groups linked by covalent bonds that are intrinsically permissible to rotations. The groups linked by such bonds are themselves comparatively rigid and constitute the fundamental dynamical elements in a protein molecule. Here we try to evaluate the importance or relevance of a specific empirical energy function to predict the folding. This function uses Ramachandran plots (Ramakrishann 1965), describing the allowed values of the dihedral angles and their populations. Here this plots are extract for a rotamer library generated from a list of 500 proteins selected form Protein Data Bank (PDB). Studies of this plots (McCammon 1984) show that they reflect the local interactions of free energy. The type of conformation are determined from a balance between local interactions (those closed to the sequence) and non-local ones. Effective statistical potentials can be extract from these populations. In quantum mechanical simulations on peptides (Bekker 1990) was found an agreement betweens' relative conformational energies and the statistical potentials. Such local potentials could be combined with non-local potential to study protein folding (Shapovalov 2011). We can expected that the rule of this local potentials is to reduce the accessible conformational space of a polypeptide by restricting the dihedral angles to values preferred by the local sequence arrangements. The use of a build-library of rotamers saves computational time producing evaluating the system energy functions, used where as the GA fitness function to determine the most likely low energy side-chain conformation.

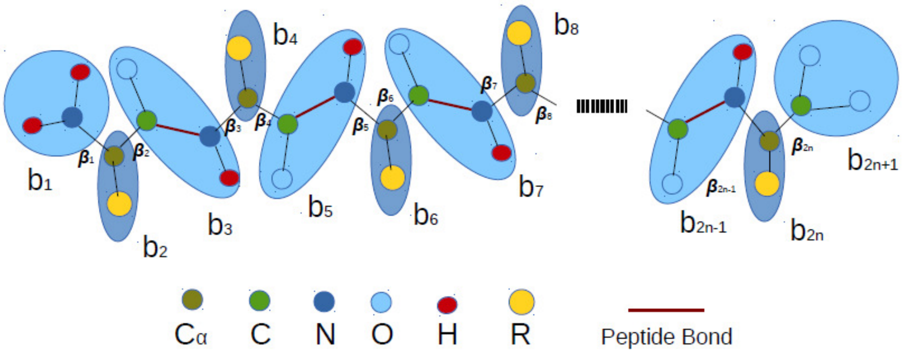


Figure 3: Refinement for a protein with n residuns and its dihedral angles.

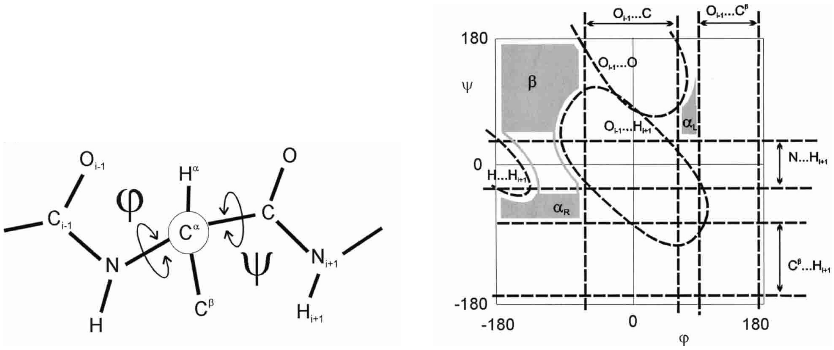


Figure 4: Fig. 1 The schematic of the alanine dipeptide that represents the protaein backbone parameterized by its dihedral angles (Bosco 2003). Fig. 2 The original Ramachandram stric map where the specific hard-sphere repulsions (dashed lines) define the allowed regions (gray) (Bosco 2003).

As described in Mandel et al. (Mandel 1977) the local interatomic distances that are directly parametrized by the dihedral angles are: $O_{i-1} \cdots C$ and $O_{i-1} \cdots C^{\beta}$ which restricts ϕ; $N \cdots H_{i+1}$ and $C^{\beta} \cdots H_{i+1}$ which restricts ψ; and $O \cdots H_{i+1}$, $H \cdots H_{i+1}$ and $O_{i-1} \cdots O$, which shaves off the corners of the allowed region (see Figure 4). It is consensual (Bosco 2003) on the basis of the conformational enumeration of polypeptide chains and molecular dynamic simulations that dihedral angles are affected by their nearest amino acid neighbours in the chain. They are affected, in particular, by the neighbour's conformation and their identity.

2 Describing a protein conformation as a 1D string and its free energy

The dihedral angles of a polypeptide chain gives a good description to the protein 3D configuration (Hoffman, 1996). The protein model as a constrained multibody system allows to described its configurations by their backbone dihedral angles. In this sense, starting with the 3D atomic coordinates of a protein molecule, its conformation can be described using a 1D sequence of dihedral angles. This simplification of protein structure is very useful on the contexts of Genetic Algorithm Optimization. It allows the direct codification, of a protein conformation, in a chromosome as a sequence of angles.

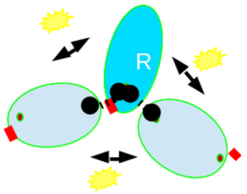
The local potential for a polypeptide chain can be computed using the probability distributions for this dihedral angles. Different fragments of this chain have very different distribution. In this work they are determined using amino acid contacts in a database of 500 known protein structures. This statistical potential are used as an iteration matrix that assigns an energy value to each possibility pair of dihedral angles. This allowed produce a low-resolution potential energy function for single residue. Here two formulas for free energy for a residuum R are tested:

$$F_1^R(\phi, \psi) = -\ln[1 + N_R(\phi, \psi)] \tag{1}$$

$$F_2^R(\phi, \psi) = -\ln[P(\phi, \psi|R)] \tag{2}$$

where $N_R(\phi, \psi)$ is the number of residues of type R, in the library, with dihedral angles ϕ, ψ (Betancourt 2004). $P(\phi, \psi|R)$ is the probability of a residuum of type R have dihedral angles ϕ, ψ. In Figure 5 we can see two iteration matrices, for two amino acids, and the associated statistical potential.

However to have sufficient information in the library about ϕ, ψ, its domain is discretized. The domain $[-\pi, \pi]$ is divided into bins of equal width. The width of a bin is determined on the expected uncertainty in the dihedral angle values (Hoffman, 1996). Here we tested different bin widths, however the bin used in obtaining must of the results presented in this document was $2\frac{\pi}{23}$, having each amino acid an iteration matrix with 23 rows and 23 columns.



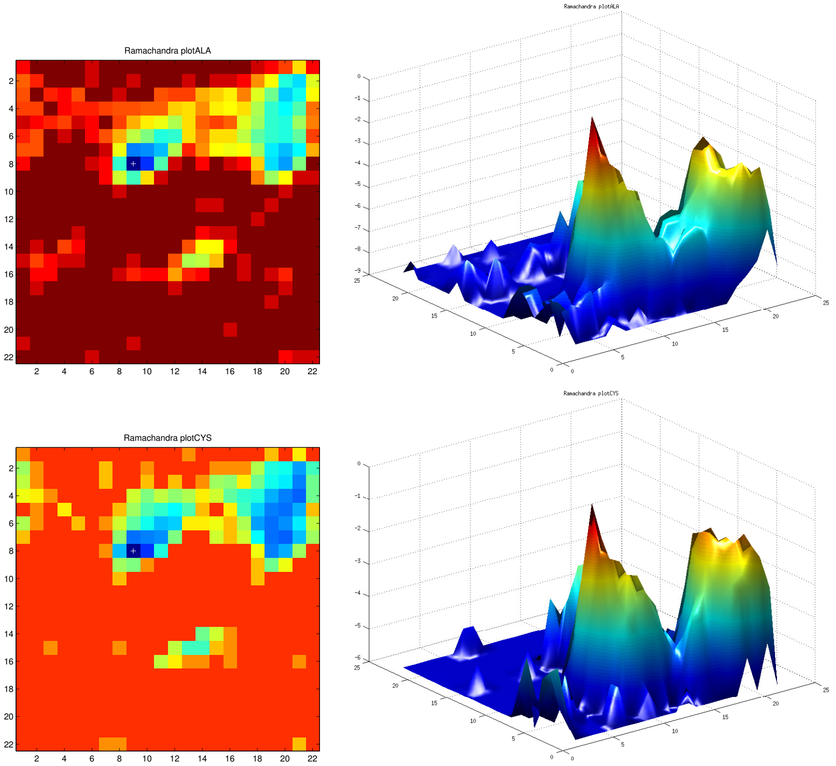


Figure 5: Local potentials are defined using Ramachandram plots. Here we present this plots for amino acids CYS and ALA. The 3D surface describes the free energy approximation using equation (2), in this cases.

Ramachandram plots or iteration matrix have been extended to the definition of statistical conformation constrains and used to compute non-local potentials. Since the conformation defined by adjacent amino acids conformations involves too many variables to be captured in a single probability density function, we divided the probability densities in individual terms involving pairs of angles. For a pair of peptide planes f_{i-1} f_i, we looked at the density plots involving (ϕ_{i-1}, ϕ_i), (ψ_{i-1}, ϕ_i), (ψ_{i-1}, ψ_i) and (ϕ_{i-1}, ψ_i). Figure 19 describe the density plots involving this angles for sequences CYS-ALA and the value of the free energy defined in equation (1) with the appropriate angle dependence.

Given a protein sequence, with length n, a dihedral angle statistical potential can be defined as in (Betancourt 2004) by

$$F_1(\Theta_k) = \sum_{i=2}^{n-1} F_1^{R_i}(\phi_i, \psi_i) + F_1^{R_{i-1}R_i}(\phi_{i-1}, \phi_i) + F_1^{R_{i-1}R_i}(\psi_{i-1}, \phi_i) F_1^{R_{i-1}R_i}(\psi_{i-1}, \psi_i) + F_1^{R_{i-1}R_i}(\phi_{i-1}, \psi_i) \tag{3}$$

where $R_{i-1}R_i$ represents a sequence of two amino acid residues and F in the sum is of the form of equation (1) with the appropriate angle dependence. A similar strategy was used to derive a formula using the free energy described by (2).

3 Protein folding perdition using GA

The basic of GA consists at least three operations - reproduction, crossover, and mutation. In this application domain a protein conformation is encoded as a sequence of dihedral angles representing a chromosome. Through the application of selection operations to initial populations of chromosomes, a new generation is formed. The initial population is here generates by the selection of random dihedral angles. The chromosomes encoding conformations with lower free energy value (our fitness function) will be kept for breeding the next generation. The next generation then is made up of copies of conformations with high fitness which form the mating pool for the following generation. The crossover operation mates a pair of chromosomes by randomly selecting crossover points and swapping the sequence parts. The nutation randomly selects a chromosomes within the population and alters part of its dihedral angles. By applying genetic operations to an initial population of protein conformations, a final population of lower energy conformations is formed.

The general procedure used here to predict the a stable conformation is the following:

1. Selection of a protein, in its native conformation, from the PDB.

2. Extraction of the protein 1° base. For a 1° base of length n, its conformations is encoded in a sequence of $2 * n - 1$ dihedral angles used as chromosome in the GA optimizer.

3. Generation of a rotamer library using probability densities of dihedral angles, one for each amino acid present in the protein, and for pairs of dihedral angles in consecutive pairs of residues in the side-chain.

4. Set the GA options.

5. Execute the GA implementation.

6. Evaluate the similarity between the predicted lower energy conformations to the protein native conformation.

In each test the quality of the predicted lower energy conformation is evaluated using mean square error (MSR) between the vector of dihedral angles, for the initial protein conformation, and vector of dihedral angles defining the produced lower energy chromosome.

For this work all results were produced using the GA Matlab implementation. Bellow it's described the selected options.

3.1 Population

Here the chromosome is defined as a vector of dihedral angles, and its length is dependent of the selected protein. Due to restriction in the available computation power the presented results were generated with initial population of 100 or 200 conformations. A best solution is to selected a population of $n \times v$ chromosomes, where b is the number of bins used in the angular discritization and v is the number of variables. For an protein defined by n residues, we have $v = 2n - 1$ variables or dihedral angles. The population type is a double vector, and the initial population is here created at random using a uniform distribution of angles in $[-\pi, \pi]$.

3.2 Fitness scaling

In the MatLab scaling function specifies the function that performs the scaling. In this work the scaling function Top is used. It scales the individuals with the highest fitness values equally.

3.3 Selection

The selection function chooses parents for the next generation based on their scaled values from the fitness scaling function. For selection function, given the problem nature the Tournament selects seems to be the best. It selects as parent protein conformation at random, and then choosing the best conformation out of that set to be a parent.

3.4 Reproduction

Reproduction options determine how the genetic algorithm creates children at each new generation. Here the Elite count is set to 5 individuals that are guaranteed to survive to the next generation. Crossover fraction specifies the fraction of the next generation that crossover produces. Mutation produces the remaining individuals in the next generation. The Crossover fraction was set in this work to 80%.

3.5 Mutation

Mutation functions make small random changes in the conformations in the population, which provide genetic diversity and enable the genetic algorithm to search a broader space. The algorithm selects a fraction of the dihedral angles for mutation, where each angle has the same probability of being mutated. Then algorithm replaces each selected angle by a random angle selected uniformly from its domains. Here a angle have a probability of 0.01 of being mutated.

3.6 Crossover

Crossover combines two protein conformations, or parents, to form a new conformation, or child, for the next generation. For crossover it was selected the Scattered method. It creates a random binary vector and selects the genes where the vector is a 1 from the first parent, and the genes where the vector is a 0 from the second parent, and combines the genes to form the child.

3.7 Stopping criteria

Stopping criteria determines what causes the algorithm to terminate. The maximum number of iterations was set to 100. If the cumulative change in the fitness function value over Stall generations is less than $1e - 8$, the algorithm stops.

I tested the *GA* optimizer with different options, sets of subpopulations and parameters, however the presented ones seems have the best performances with short proteins.

4 Experiments

To discuss the searching ability of GA for the minimizing the empirical energy functions 1 and 2, a set of small proteins are targeted. The predicted configuration is compared with its native conformation.

The selected proteins are:

1. 1PLX - Met-enkephain 1 - 1 chain - 5 amino acids
2. 1NOT - GI alpha conotoxin - 1 chain - 14 amino acids
3. 1PEN - Alpha-conotoxin pnia - 1 chain - 17 amino acids
4. 2FKL - Amyloid beta A4 protein predursor - 2 chain - 66 amino acids

4.1 1PLX - Met-enkephain 1

Enkephalins are pentapeptides found in the central nervous system, its native conformation can be seen in Figure 6.

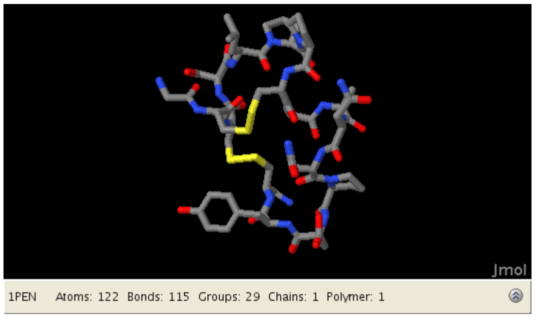


Figure 6: 1PLX

It is believed that these neuropeptides interact with the nerve cell membrane to adopt a conformation suitable for their binding to an opiate receptor. It is defined by one chain of 5 amino acids, TYR-GLY-GLY-PHE-MET, having its conformation described by 8 dihedral angles. In its native conformation the empirical energy given by (1) is -28.6400 and by formula (2) is 38.3750.

Table 1 presents the MSE for each pair of dihedral angles in a peptide plane and residuum. This predicted conformation was computed using the energy function defined by formula 1. We can compare the prediction and the stable conformation on Figure 8. Figure 7 presents the bins for native dihedral angles, the predicted bins for dihedral angles, the for each angle, and the error histogram. For this empirical energy the predicted conformation have a MSE=1.22.

contact	n. residues	MSE
TYR-GLY	1	9.055
GLY	2	1.500
GLY-GLY	1	2.828
GLY-PHE	1	1.000
PHE	1	1.000
PHE-MET	1	2.000

Table 1: 1PLX with energy function (2): Number of pairs of dihedral angle pair in petind planes and residues and its MSE.

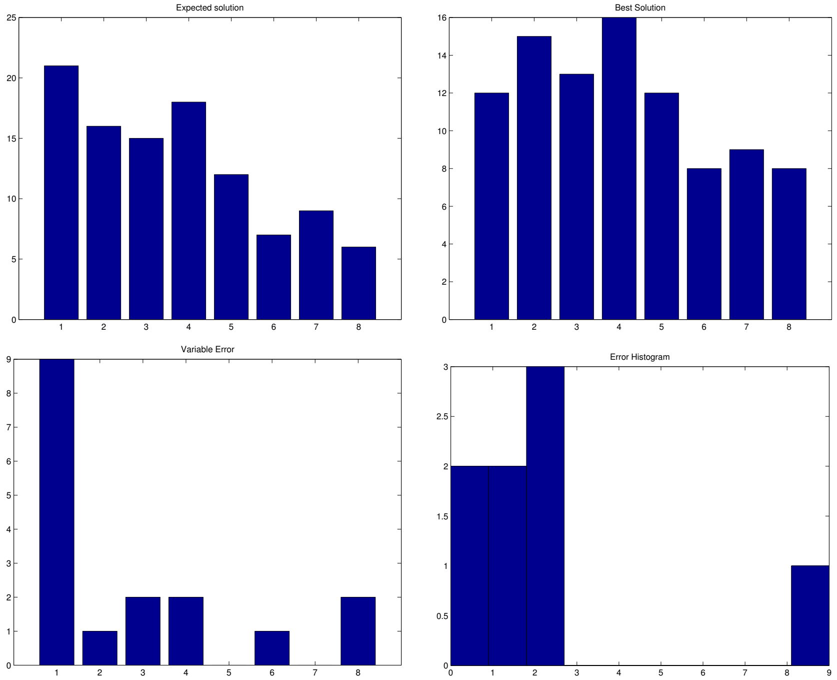


Figure 7: 1PLX:Using energy given by formula 2 and a bin bensity of: 0.3 Expected free energy: 38.375 Execution time: 131.92 Predicted energy: 16.16 MSE 1.22. Image 1 - Exact dihedral angles for 1PLX. Image 2 - Predicted dihedral angles. Iamge 3 - Error in each angle. Image 4 - Error histogram.

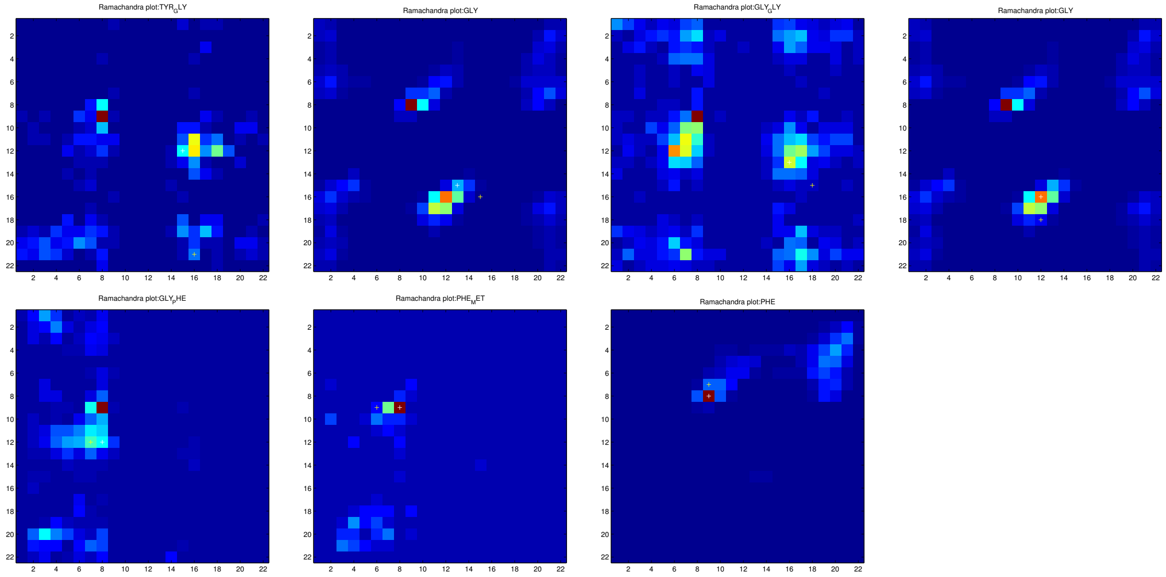


Figure 8: 1PLX density distributoins, with the stable dihedral angles (yellow) and the predicted dihedral angle (white) using formula 2.

Table 3 presents the MSE for each pair of dihedral angles. This predicted conformation was computed using the energy function defined by formula 1. Figure 9 described in bar graphs the bins for native dihedral angles, the predicted bins for dihedral angles, the for each angle, and the error histogram. For this empirical energy the predicted conformation have a MSE=1.82.

contact	cases	MSE
TYR-GLY	1	9.000
GLY	2	1.581
GLY-GLY	1	3.162
GLY-PHE	1	0.000
PHE	1	11.000
PHE-MET	1	11.045

Table 2: 1PLX with energy function (1): Number of pairs of dihedral angle pair each petind plane and residuum type and its MSE to the native conformation.

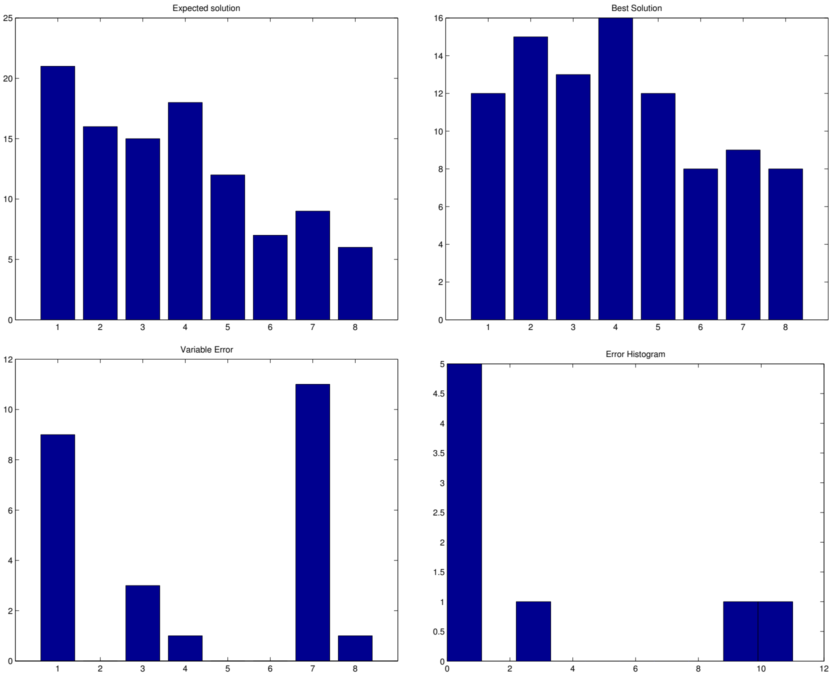


Figure 9: 1PLX: Using formula 1 and a bin density of: 0.3 Expected free energy: -28.6400 Execution time: 98.71 Predicted energy: -47.41 MSE 1.82. Image 1 - Exact dihedral angles for 1PLX. Image 2 - Predicted dihedral angles. Iamge 3 - Error in each angle. Image 4 - Error histogram.

4.2 1NOT GI - alpha conotoxin

Predatory marine snails of the genus Conus paralyze their fish prey by injecting a potent toxin. The alpha-conotoxin GI is a 13-residue peptide isolated from venom of Conus geographus. Its native structure can be seen in Figure 10

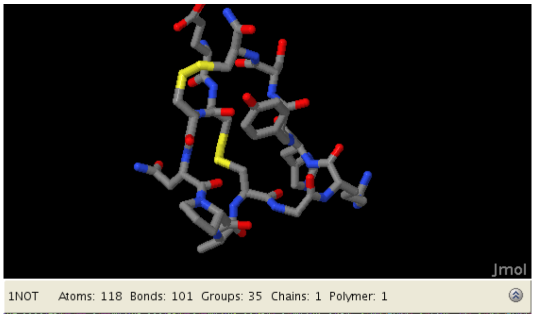


Figure 10: 1NOT

The following Table presents the MSE for each pair of dihedral angles. This predicted conformation was computed using the energy function defined by formula 2. Figure 11 presents the bins for native dihedral angles, the predicted bins for dihedral angles, the error in each angle, and the error histogram. For this empirical energy the predicted conformation have a MSE=0.73. When the GA optimizer uses empirical energy (1) the predicted conformation have a MSE of 1.27 and the results of this can be seen in Figure 15.

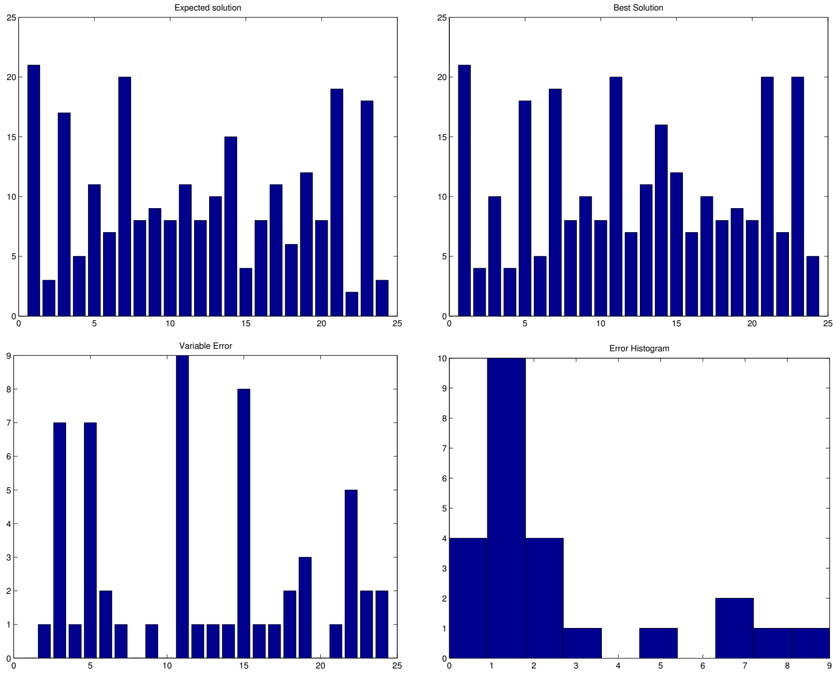


Figure 11: 1NOT: Num. Var.: 24 Density: 0.3. Expected free energy: 114.71 Execution time: 231.66 Predicted energy 82.18 . MSE 0.73. Image 1 - Exact dihedral angles for 1NOT. Image 2 - Predicted dihedral angles. Iamge 3 - Error in each angle. Image 4 - Error histogram.

Residuum	Cases	MSE
CYS	3	3.367
ASN	1	2.236
PRO	1	1.000
ALA	1	9.000
GLY	1	8.062
ARG	1	1.414
HIS	1	3.606
TYR	1	1.000
SER	1	5.385

Table 3: 1PLX with energy function (1): Number of pairs of dihedral angles for each residuum type and its MSE to the native conformation.

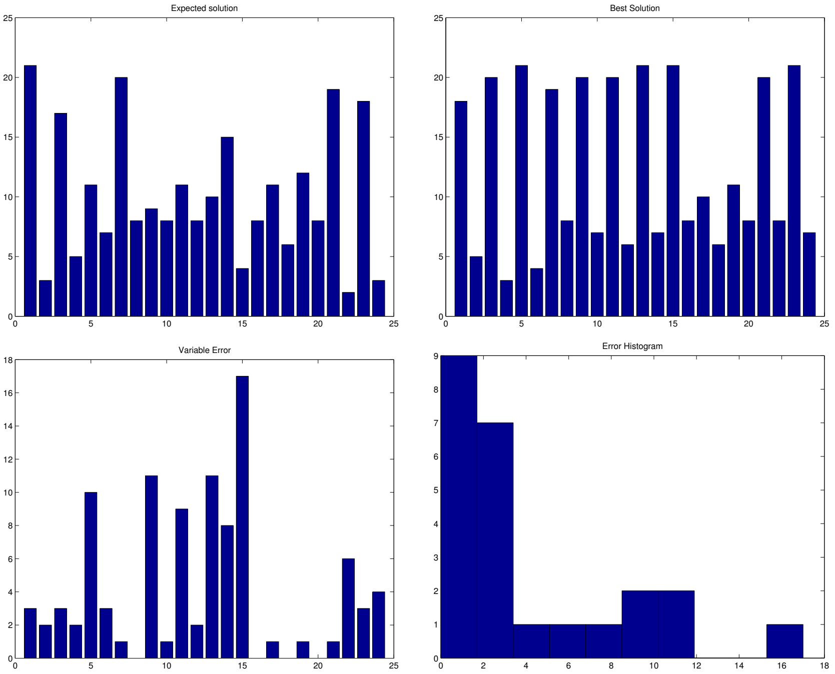


Figure 12: 1NOT: Num. Var.: 24 Density: 0.3. Expected free energy: -100.86 Execution time: 223.46 Perdicted energy -127.08 . mse 1.24. Image 1 - Exact dihedral angles for 1NOT. Image 2 - Predicted dihedral angles. Iamge 3 - Error in each angle. Image 4 - Error histogram.

4.3 1PEN - Alpha-conotoxin pnia

Alpha-Conotoxins are peptide toxins, isolated from Conus snails, that block the nicotinic acetylcholine receptor (nAChR) Its native structure is presented in Figure 13. It is defined by 1 chain of 17 amino acids.

Residuum	Cases	MSE
CYS	3	3.887
SER	1	2.828
LEU	1	9.487
PRO	3	5.228
ALA	2	0.000
ASN	2	2.179
ASP	1	0.000
TYR	1	2.000

Table 4: 1PEN with energy function (2): Number of pairs of dihedral angles for each residuum type and its MSE to the native conformation.

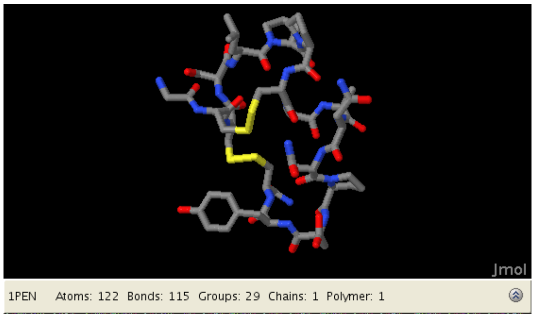


Figure 13: 1PEN

Table 4 presents the MSE for each pair of dihedral angles. This predicted conformation was computed using the energy function defined by formula (2). Figure 14 presents the bins for native dihedral angles, the predicted bins for dihedral angles, the error in each angle, and the error histogram. For this empirical energy the predicted conformation have a MSE=0.76. When the GA optimizer uses empirical energy (1) the predicted conformation have a MSE of 0.79 and the results of this can be seen in Figure 15.

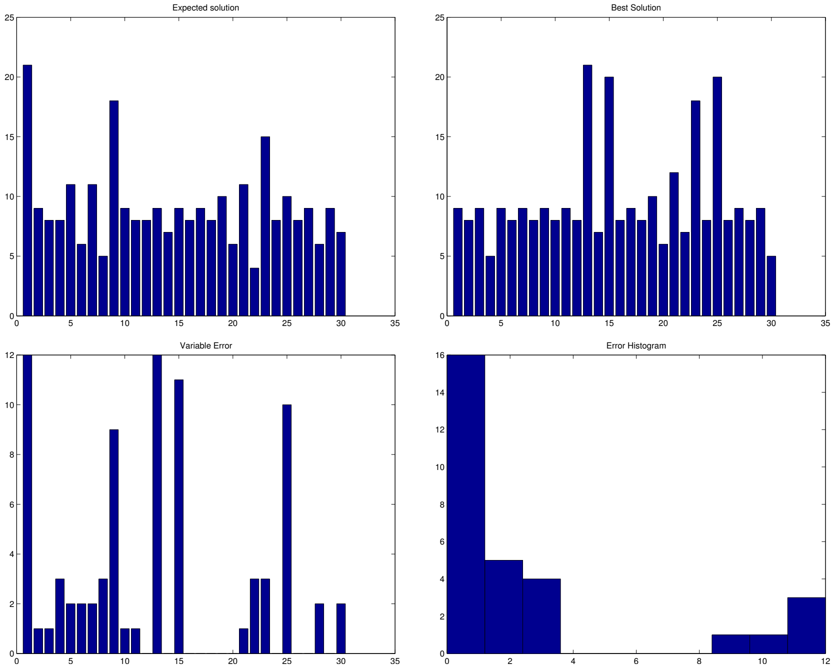


Figure 14: 1PEN: Using formula 2 and a density of 0.3 Expected free energy: 119.7001 Execution time: 286.15 Predicted energy: 73.49 MSE 0.85. Image 1 - Exact dihedral angles for 1PEN. Image 2 - Predicted dihedral angles. Iamge 3 - Error in each angle. Image 4 - Error histogram.

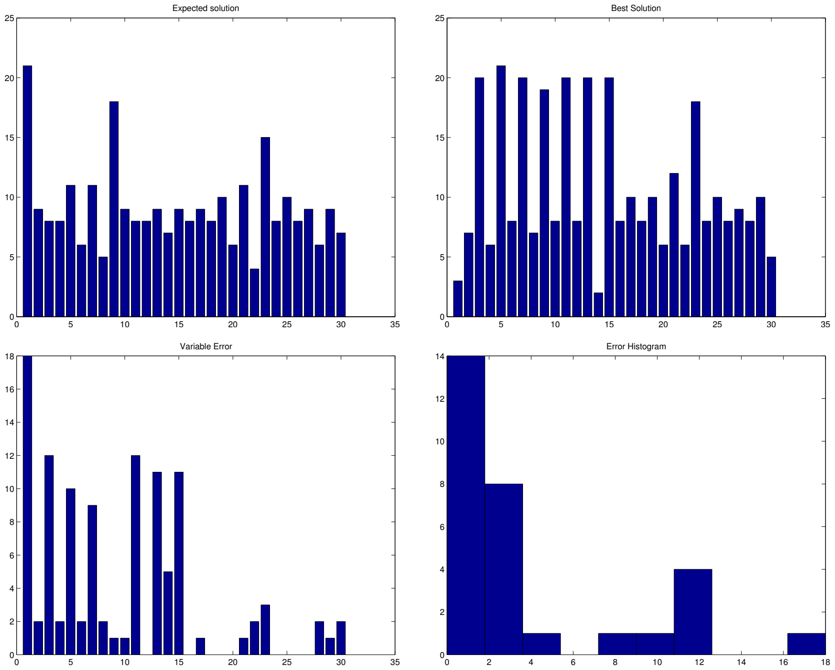
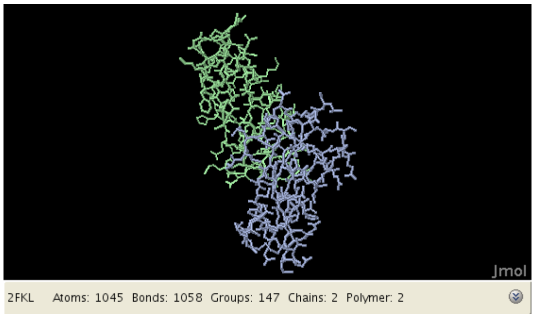


Figure 15: 1PEN: Using formula 1 and a density of 0.3 Expected free energy: -177.86 Execution time: 284.12 Predicted energy: -211.97 MSE 1.11. Image 1 - Exact dihedral angles for 1PEN. Image 2 - Predicted dihedral angles. Iamge 3 - Error in each angle. Image 4 - Error histogram.

4.4 2FKL - Amyloid beta A4 protein predursor

2FKL is a 2 chain protein with 66 amino acids. Its structure is presented in Figure 16. Here we apply the GA optimization to predicted a stable conformation only for its first chain.



2FKL Atoms: 1045 Bonds: 1058 Groups: 147 Chains: 2 Polymer: 2

Figure 16: 2FKL

Table 5 presents the MSE for each pair of dihedral angles. This predicted conformation was computed using the energy function defined by formula 2. Figure 17 presents the bins for native dihedral angles, the GA predicted bins for each dihedral angles, the error in each angle, and the error histogram. For this empirical energy the predicted conformation have a MSE=0.57. When the GA optimizer uses empirical energy (1) the predicted conformation have a MSE of 0.79 and the results of this can be seen in Figure 18.

Residuum	num	mse
LEU	7	2.470
VAL	5	2.514
PRO	3	5.153
ASP	4	7.062
LYS	5	2.987
CYS	6	3.167
PHE	3	4.595
HIS	5	2.966
GLN	1	9.487
GLU	5	3.768
ARG	2	10.025
MET	2	4.975
THR	4	2.646
TRP	1	2.236
ALA	1	11.000
SER	2	5.431
ASN	1	1.000
TYR	1	9.000
GLY	3	12.083
ILE	1	1.000

Table 5: 2FKL with energy function (2): Number of pairs of dihedral angles for each residuum type and its MSE to the native conformation.

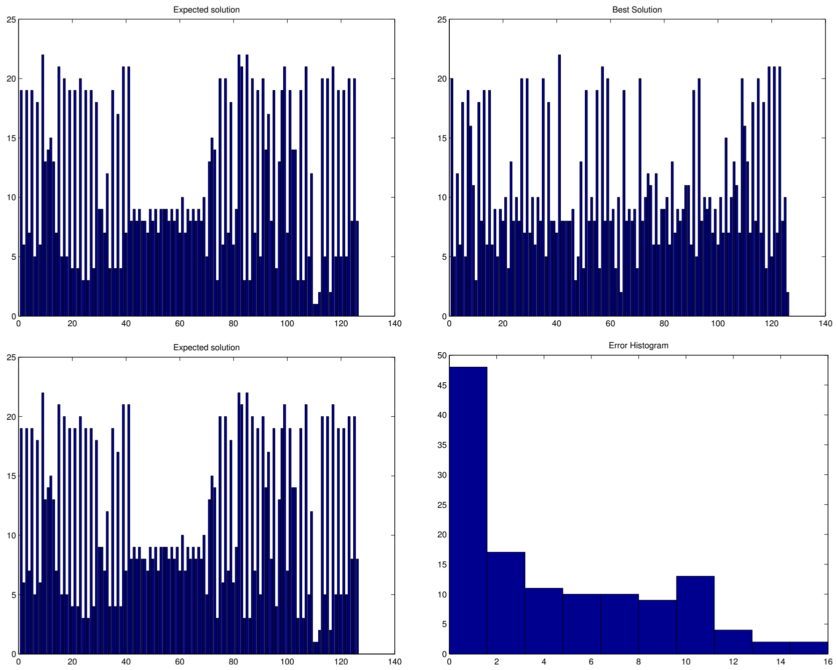


Figure 17: 2FKL: : Using formula (2) and density: 0.3. Expected free energy: 727.65. Execution time: 3039.49. Predicted energy 739.28. L2 error 0.54. Image 1 - Exact dihedral angles for 2FKL. Image 2 - Predicted dihedral angles. Iamge 3 - Error in each angle. Image 4 - Error histogram.

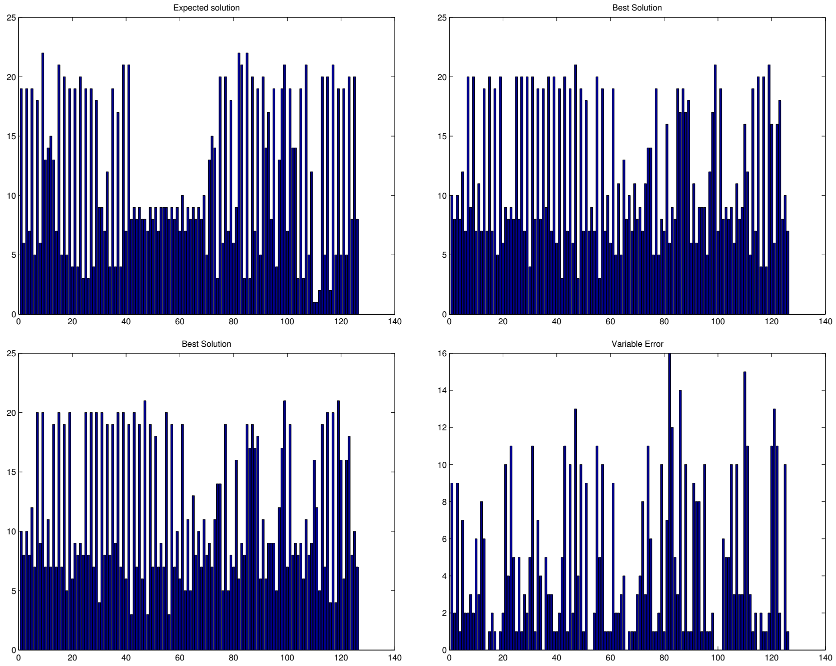


Figure 18: 2FKL:: Using formula (1) and a density of 0.3. Expected free energy: -642.6380. Execution time: 3970.66. Predicted energy -685.38. L2 error 0.54. Image 1 - Exact dihedral angles for 2FKL. Image 2 - Predicted dihedral angles. Iamge 3 - Error in each angle. Image 4 - Error histogram.

5 Conclusions

From the presented results we can concluded that the presented methodology have an inefficient description of the dynamics of short proteins. This can be the result of using dihedral angle probability distributions in a window of two consecutive amino acids. To improve it I should impose stronger restrictions to the protein kinematic. I expect to have the opportunity to test the above methodology with statistical potentials extracted from rotames libraries for sequences of three amino acids as proposed by Betancourt, J. Skolnick in(Betancourt 2004). However the empirical energy formulate in 1 and used in (Betancourt 2004), produced a worst prediction performance, for short proteins, than the probabilistic energy formulation give in 2. In the future I intended continue the testing with a test set of bigger proteins, and extend the predictions for conformations in more than one chain.

REFERENCES

(Altis 2008) A. Altis, *Modeling the Free Energy Landscape of Biomolecules via Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulation* , Thesis Goethe-Universitat, 2008.

(Bekker 1990) H. Bekker, *Molecular Dynamics Simulation Methods Revised*, Thesis Rijksuniversiteit Groningen, 1990.

(Betancourt 2004) M. Betancourt, J. Skolnick, *Local Propensities and Statistical Potentials of Backbone Dihedral Angles in Proteins*, J. Mol. Biol. 342, 635-649, 2004

(Bosco 2003) H. Bosco, A. Thomas, R. Brasseur, *Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in α-helix*, Protein Science, 12:2508-2522, 2003.

(Hoffman, 1996) D. Hoffman, *Comparing of protein Structures by Transformation into Dihedral Angle Sequences*,Phd theses University of North Carolina, 1996.

(Karlson 2012) K. Karlson *Multiscale Continuum Modeling of Protein Dynamics*, Georgia Institute of Technology, 2012.

(Mandel 1977) N. Mandel, G. Mandel, B. Trus, J.Carlson, R. Dickerson, *Tuna cytochrome c at 2 A resolution*, Coordinate optimization and comparison of structures. J. Biol. Chem. 252: 4619-4636, 1977.

(McCammon 1984) J. McCammon, *Protein dynamics*, Rep. Prog. Phys., Vol 47, pp 1-46, 1984.

(Ramakrishann 1965) C. Ramakrishnan, G. Ramachandran, *Stereochemical Criteria for Polypeptide and Protein Chain Conformation*, Biophys J. 5(6), pp 909-933.

(Shapovalov 2011) M. Shapovalov, R. Dunbrack, *A Smothed Backbone-Dependent Rotamer Library for Protains Derived from Adaptive Kernel Density Estimates and Regression*, Structure, 19, pp. 844-856, 2011.

Marcotte, I., Separovic, F., Auger, M., Gagne, S.M. *A multidimensional 1H NMR investigation of the conformation of methionine-enkephalin in fast-tumbling bicelles.* Biophys.J. 86: 1587-1600, (2004).

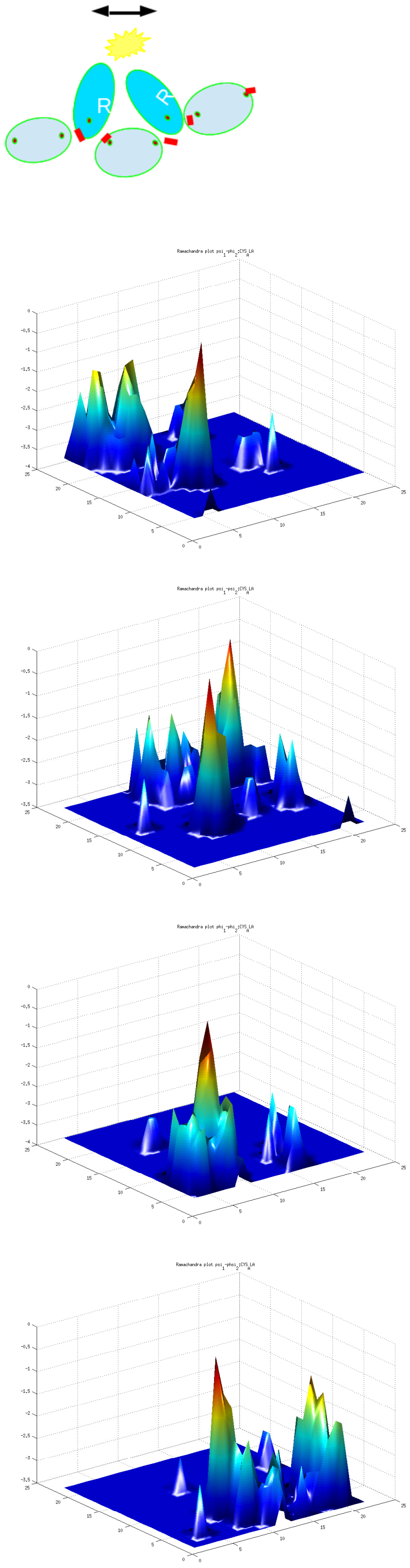


Figure 19: Density plots involving sequences CYS-ALA and the value of the free energy defined in equation 1 with the appropriate angle dependence

6 Apendix

The presented results were generated using 2 scripts in Python, 1 MatLab script and 1 Matlab function (https://github.com/MelroLeandro/FoldingWIthRotamerAndGA.git):

1. Extract_tests.py,is locate in the folder "./test" and it is used to extract the protein 1^o base and dihedral angles from a pdb file format.

2. ramachandran_Res.py, is located in the folder "./top500" and it is used to generated information used in Ramachandran plots construction.

3. Protein_stable_multistate.m is the script used to predict protein folding using GA. This scripts reads a 1^o base sequences and dihedral angles from folder "." and uses data in the folder "./top500" for density plots generation.

4. ObjFun.m is a MatLab function used by Protein_stable_multistate.m and the GA optimizer to compute the empirical energy surface.

The Python scrips have dependences. Here libraries numpy, scipy, matplotlib and BioPython are used.

The results and the auxiliary data are distributed by tree folders . , ./test and ./top500.

1. The folder . is the location where folding prediction graphs are saved, and where the proteins 1^o base file and the file with information about the dihedral data must be located.

2. The folder ./top500 is the location for 500 proteins, in pdf format, and it location of rotamer library, for each amino acid and ech sequence of two amino acids, defined by files in text format, with respectively two and four columns.

A protein folding prediction is generated by executing the Matlab script Protein_stable_multistate.m, here we may parametrise the solver, select the protein, select the empirical energy, configure the execution. Bellow you can see the top of this script, and test it to predict the native configuration for protein '1FXD'.

```
%%
% Data selection
%    You must selected here the protein

%pdb_code='2FKL';
%pdb_code='1PEN';
%pdb_code='1NOT';
%pdb_code='1AIE';
%pdb_code='1AJJ';
pdb_code='1FXD'; % <--- my selection
%pdb_code='1PLX';
%%
% Define discritization and ga population
%
density=0.3;   % degree of dicretization in the Ramachandran plot

% Selection 1 if you want dynamic constrains imposed
% by density plots for sequences of two amino acids

constrains = 1;  % 1 statistical constrain
                 % 0 without constrains
% Selection of your empirical energy function
%
prob=1;          % 1 uses the probabilistic energy
                 % 0 uses the statistical energy

% Number of random conformations for GA
%
PopulationSize_Data = 200 % GA initial population

% Different conformation initializations
%
starting_population = 0; % Criteria to strat population
                         % 0 - [-pi,pi] random
                         % 1 - uses expected solution
                         % 2 - uses dihedral angles on each amino acid
                         % 3 - uses contact dihedral angles

%%
% Selection of graphic output
draw_structure=1;  % Display and manipulate 3D molecular structure
                   % 1 - on
                   % 0 - off

plot_R = 0;     % Draw Ramachandran plots
                % 1- 2D plot
                % 2- 3D plot

check_prot = 0; % check dihedral angles for expected solution
                % on Ramachandran plot

save_img = 1;   % Save graphics in eps format
                % 1 - on
                % 0 - off

plot_Solution = 1; % Display the Ramachandran plot with the
                   % predictions and native values.
```

After execute the script: It begins by showing the protein 3D structure. Press ENTER to save the best view of its structure.

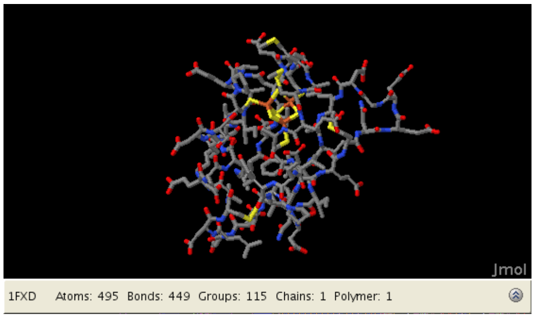


Figure 20: 1FXD

Then the scripts start the construction of density plots to be stored in global variables.

```
Computing  GLU probability densities...
Computing  GLU_ALA probability densities...
Computing  ALA probability densities...
Computing  ALA_CYS probability densities...
Computing  CYS probability densities...
Computing  CYS_VAL probability densities...
Computing  VAL probability densities...
Computing  VAL_GLU probability densities...
Computing  GLU_ILE probability densities...
Computing  ILE probability densities...
Computing  ILE_CYS probability densities...
Computing  CYS_PRO probability densities...
Computing  PRO probability densities...
Computing  PRO_ASP probability densities...
Computing  ASP probability densities...
Computing  ASP_VAL probability densities...
Computing  VAL_PHE probability densities...
Computing  PHE probability densities...
Computing  PHE_GLU probability densities...
Computing  GLU_MET probability densities...
Computing  MET probability densities...
Computing  MET_ASN probability densities...
Computing  ASN probability densities...
Computing  ASN_GLU probability densities...
Computing  GLU_GLU probability densities...
Computing  GLU_GLY probability densities...
Computing  GLY probability densities...
Computing  GLY_ASP probability densities...
Computing  ASP_LYS probability densities...
Computing  LYS probability densities...
Computing  LYS_ALA probability densities...
Computing  ALA_VAL probability densities...
Computing  VAL_VAL probability densities...
Computing  VAL_ILE probability densities...
Computing  ILE_ASN probability densities...
Computing  ASN_PRO probability densities...
Computing  ASP_SER probability densities...
Computing  SER probability densities...
Computing  SER_ASP probability densities...
Computing  ASP_LEU probability densities...
Computing  LEU probability densities...
Computing  LEU_ASP probability densities...
Computing  ASP_CYS probability densities...
Computing  ALA_ILE probability densities...
Computing  ILE_ASP probability densities...
Computing  SER_CYS probability densities...
Computing  PRO_ALA probability densities...
Computing  ALA_GLU probability densities...
Computing  ILE_VAL probability densities...
Computing  VAL_ARG probability densities...
Computing  ARG probability densities...
Computing  ARG_SER probability densities...
---------------------------------
- Starting optimal conformation
-
-   Protein: 1FXD
-   Population: 200
-   Num. Var.: 92
-   Num. of stable solutions: 1
-   Expected free energy: 441.7794
---------------------------------
```

Then it starts the GA optimizer.

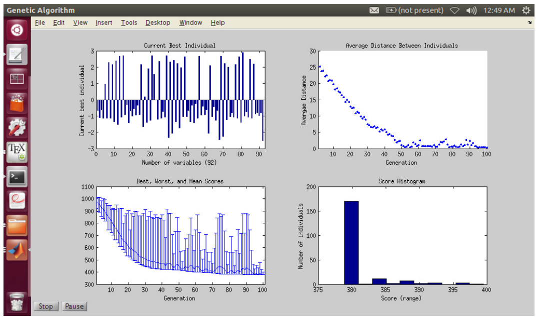


Figure 21: 1PEN

When the GA finish its task, it is plotted a bar with the exact values for each dihedral angles on 1FXD, a bar with the values of predicted dihedral angles, a bar with the error in each angle and the error histogram. It also display a table with the error in each contact, in our case:

```
Time 4348.39
Energy   378.91
Min. L2 error for stable conformation 1: 0.56
------- Error by Att --------------
|  Att: GLU_ALA  num: 3   error: 2.728
|  Att: ALA   num: 5   error: 3.736
|  Att: ALA_CYS  num: 1   error: 0.000
|  Att: CYS   num: 4   error: 2.031
|  Att: CYS_VAL  num: 2   error: 3.500
|  Att: VAL   num: 6   error: 3.193
|  Att: VAL_GLU  num: 2   error: 6.745
|  Att: GLU   num: 7   error: 3.251
|  Att: GLU_ILE  num: 1   error: 10.000
|  Att: ILE   num: 4   error: 4.337
|  Att: ILE_CYS  num: 1   error: 10.440
|  Att: CYS_PRO  num: 2   error: 2.000
|  Att: PRO   num: 3   error: 5.598
|  Att: PRO_ASP  num: 2   error: 6.910
|  Att: ASP   num: 6   error: 2.034
|  Att: ASP_VAL  num: 1   error: 2.236
|  Att: VAL_PHE  num: 1   error: 4.123
|  Att: PHE   num: 1   error: 10.770
|  Att: PHE_GLU  num: 1   error: 10.770
|  Att: GLU_MET  num: 1   error: 11.180
|  Att: MET  num: 1   error: 2.236
|  Att: MET_ASN  num: 1   error: 1.414
|  Att: ASN  num: 2   error: 5.268
|  Att: ASN_GLU  num: 1   error: 10.050
|  Att: GLU_GLU  num: 2   error: 4.583
|  Att: GLU_GLY  num: 1   error: 11.045
|  Att: GLY  num: 1   error: 2.236
|  Att: GLY_ASP  num: 1   error: 3.606
|  Att: ASP_LYS  num: 1   error: 9.220
|  Att: LYS  num: 1   error: 12.530
|  Att: LYS_ALA  num: 1   error: 11.045
|  Att: ALA_VAL  num: 1   error: 0.000
|  Att: VAL_VAL  num: 1   error: 1.414
|  Att: VAL_ILE  num: 1   error: 1.000
|  Att: ILE_ASN  num: 1   error: 9.487
|  Att: ASN_PRO  num: 1   error: 1.414
|  Att: ASP_SER  num: 2   error: 4.416
|  Att: SER  num: 2   error: 5.220
|  Att: SER_ASP  num: 1   error: 10.198
|  Att: ASP_LEU  num: 1   error: 1.000
|  Att: LEU  num: 1   error: 0.000
|  Att: LEU_ASP  num: 1   error: 1.000
|  Att: ASP_CYS  num: 1   error: 0.000
|  Att: ALA_ILE  num: 2   error: 7.762
|  Att: ILE_ASP  num: 1   error: 10.050
|  Att: SER_CYS  num: 1   error: 2.236
|  Att: PRO_ALA  num: 1   error: 10.050
|  Att: ALA_GLU  num: 1   error: 12.207
|  Att: ILE_VAL  num: 1   error: 5.000
|  Att: VAL_ARG  num: 1   error: 12.369
|  Att: ARG  num: 1   error: 9.487
|  Att: ARG_SER  num: 1   error: 9.487
-----------------------------------
```

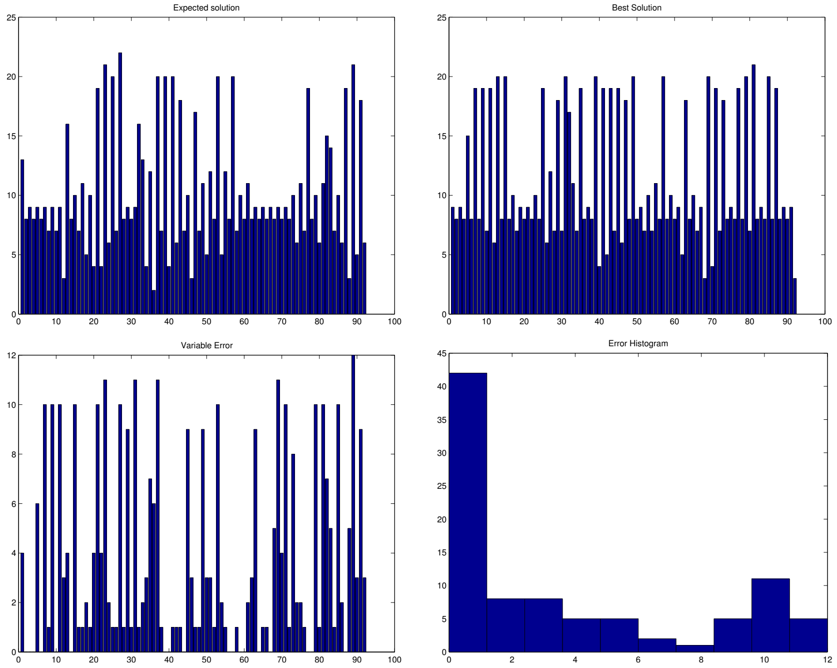


Figure 22: 1FXD: Using formula 2 and a density of 0.3 Expected free energy: 441.7794 Execution time: 4348.39 Predicted energy: 378.91 MSE 0.56. Image 1 - Exact dihedral angles for 1PEN. Image 2 - Predicted dihedral angles. Image 3 - Error in each angle. Image 4 - Error histogram.

After this you can see the plots with the exact values and the predicted values for each dihedral pair in the correspondent density plots.

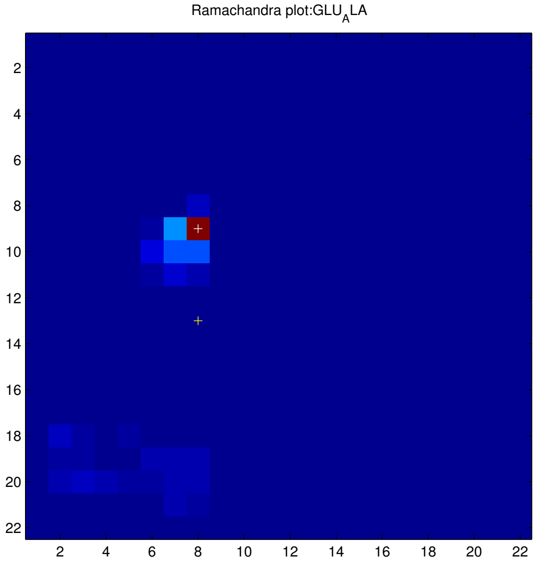


Figure 23: 1FXD: Density plot, the exact values and the predicted values for a dihedral pair