A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides

Márcio Dorn[1] Mario Inostroza-Ponta[2] Luciana S. Buriol[1]
Hugo Verli[3]

[1]Institute of Informatics, UFRGS, Porto Alegre, Brazil

[2]Departamento de Ingeniería Informática, UdeSantiago, Santiago, Chile

[3]Center for Biotechnology, UFRGS, Porto Alegre, Brazil

CEC2013, Cancún

Outline

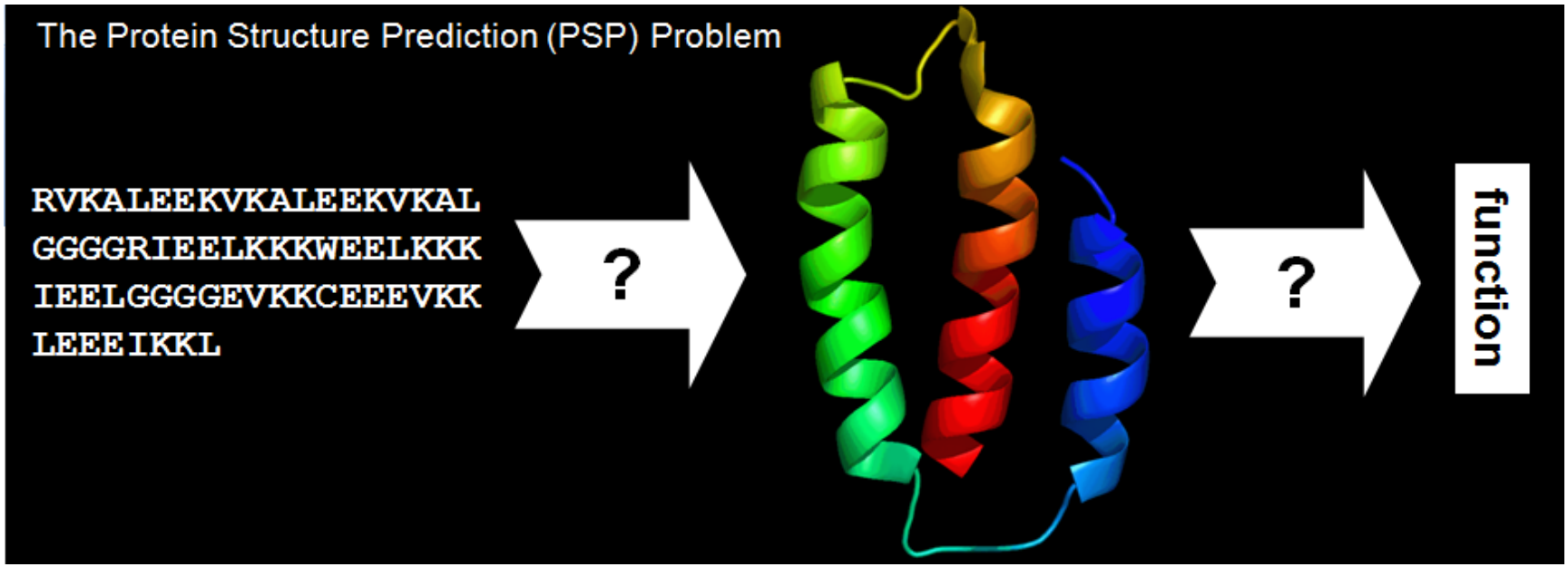
Motivation

Motivation

- Biological data explosion in the mid 1990s
- According to Protein Data Bank:
  - ∼164 million "protein" sequences (GeneBank), ∼6 million are non-redundant sequences (NR)
- Number of 3D structures in the PDB (on 8th of May, 2013)
  - 84,768 protein 3D structures
  - 1,393 distinct folds
- Clearly, there is a **gap** between the **number of protein sequences generated** and the number of new **protein folds determined** by experimental methods such as X-ray diffraction and NMR.

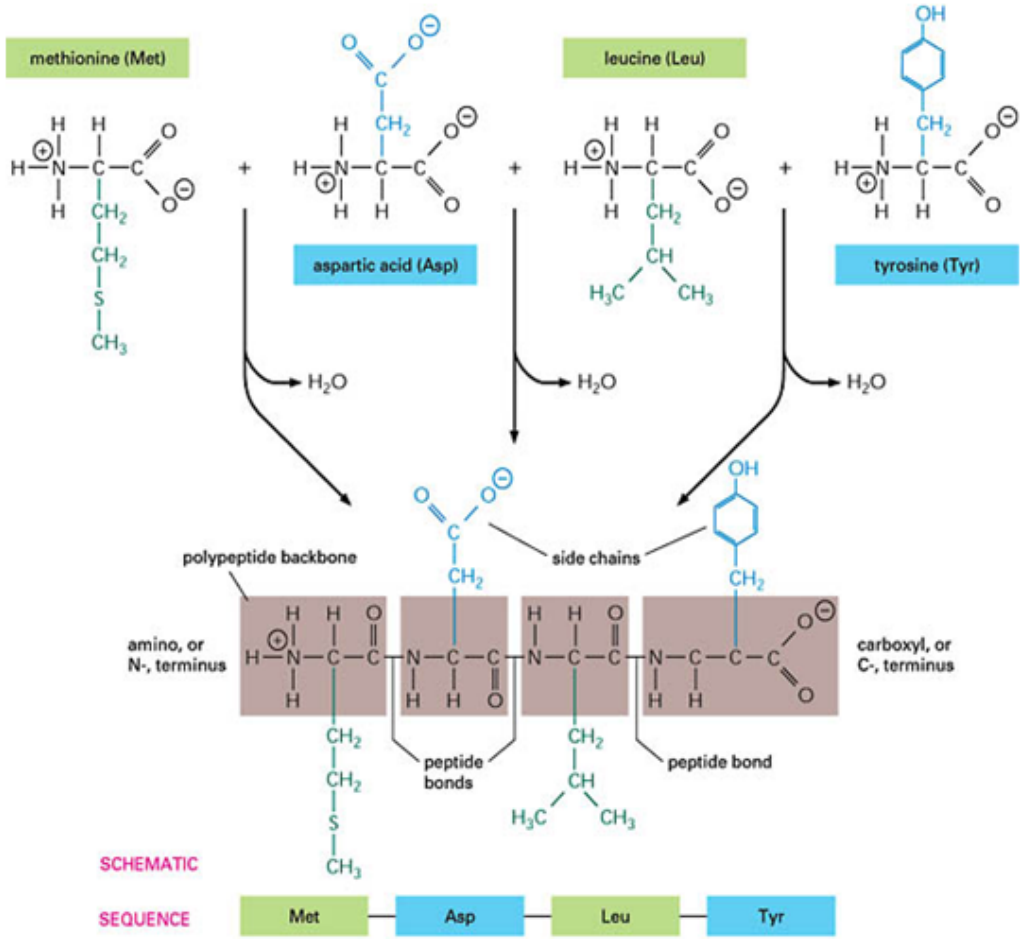
Proteins and Polypeptides

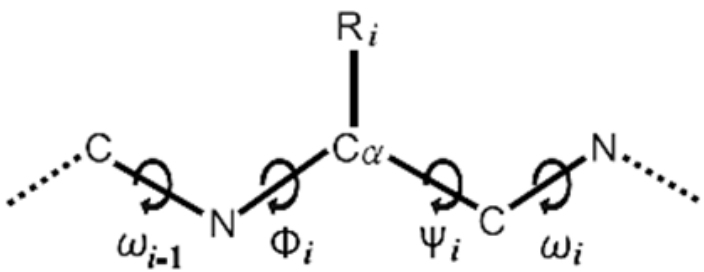
- Proteins play a variety of functions on the cell: structural, catalysis in chemical reactions, transport and storage, regulatory proteins, recognition control, among others;
- From a structural point of view, a protein is an ordered linear chain of building blocks known as **amino acid residues**;
- There are 20 different **amino acid residues**. Each amino acid is composed by:
  - an amino group, a carboxyl group, and a variable side chain, bond to an α-carbon
- The activity or function of the protein is governed by its three-dimensional structure.



- 3D PSP was shown to belong to NP-Complete class.
- Goal is **to predict the native structure of a protein molecule**.

Proteins: Peptide Bond





Knowledge-based approach proposal

Using the knowledge stored in PDB

In theory torsion angles ϕ and ψ can take values between -180 and 180.

- Protein Data Bank (*http://www.rcsb.org*)
  - Structures of the ∼80,000 proteins determined experimentally (crystallography or NMR spectroscopy)
  - Angles ϕ and ψ are known
  - Creates a rich source of information about angles

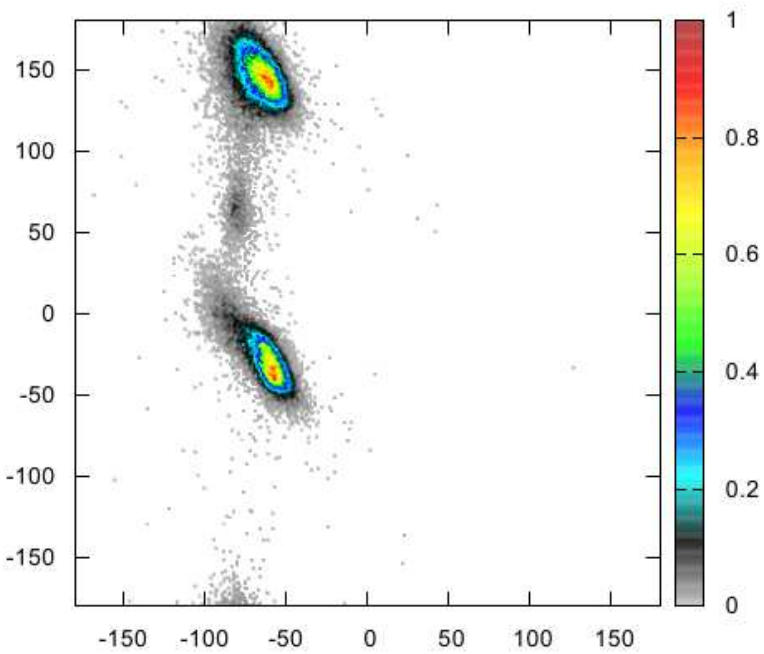
Using the knowledge stored in PDB

In theory torsion angles ϕ and ψ can take values between -180 and 180.

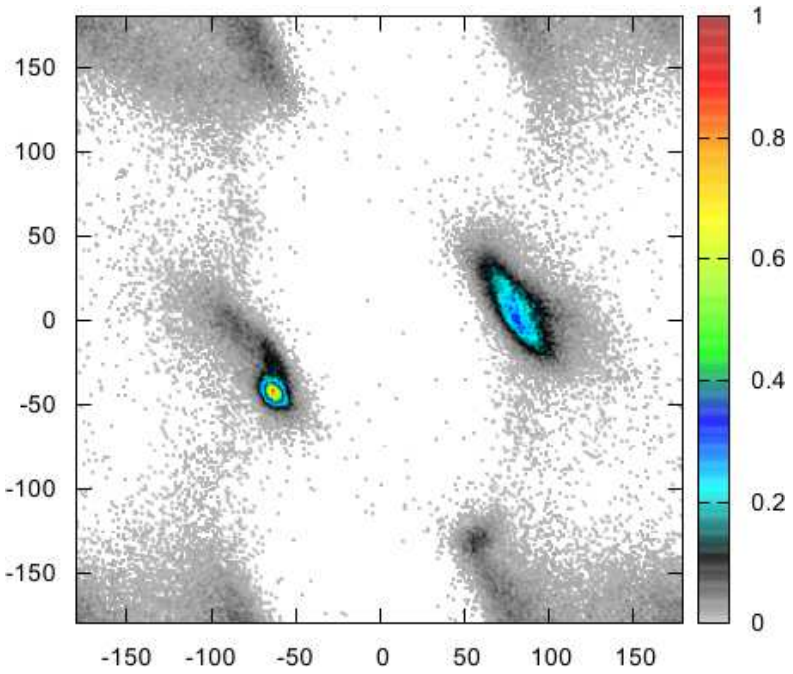
- Protein Data Bank (*http://www.rcsb.org*)
  - Structures of the ∼80,000 proteins determined experimentally (crystallography or NMR spectroscopy)
  - Angles ϕ and ψ are known
  - Creates a rich source of information about angles

Histogram plot

Proline



Glycine



Local Search strategy

- We build a histogram of the main-chain torsion angles matrix for each amino acid
  - For each pair of torsion angles $(\phi, \psi)(i \leq \phi < i+1, j \leq \psi < j+1)$ we increase H(i,j) by one
- To increase more dense regions:

$$H'_a(i,j) = \sum_{r=i-1}^{i+1} \sum_{s=j-1}^{j+1} H_a(r,s) \tag{1}$$

- We compute the probability of the pairs of angles to be in a cell (i,j) by

$$AP_a(i,j) = \frac{H'_a(i,j)}{\sum_{\forall x,y} H'_a(x,y)} \tag{2}$$

- We build a list of Angles Probabilities for each amino acid a named APL_a
- **We use APL to search for more probable angles combination and combined it with a Genetic Algorithm**

General schema

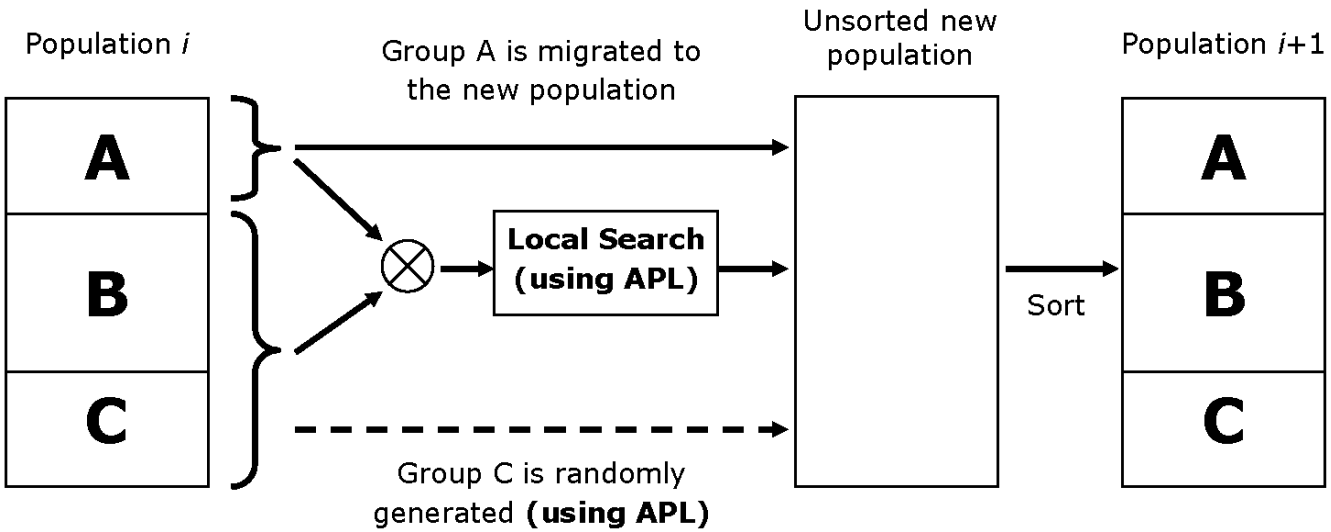


Figure: Schema of one iteration of GA used for the PSP problem. The *APL* is also used in the generation of the initial population.

Algorithm 1 GA with Local Search for the 3-D PSP Problem

Input: A protein given as a sequence of amino acids;
$Pop^0 \leftarrow$ **Generate initial population using** APL;
Sort individuals and define groups A, B and C;
for $i = 1$ to $NGen$ **do**
    $Pop^i(A) \leftarrow Pop^{i-1}(A)$
    for $j = 1$ to $|B|$ **do**
        $P_1 \leftarrow getIndividual(A)$;
        $P_2 \leftarrow getIndividual(B + C)$;
        $Offspring \leftarrow$ **Crossover**(P_1, P_2)
        $Offspring \leftarrow$ **LocalSearch**$(Offspring)$
        $add(OffSpring, Pop^i(B))$
    end for
    for $j = 1$ to $|C|$ **do**
        $Pop^i(C) \leftarrow$ **Generate Individual using** APL
    end for
    sort(Pop^i), $best \leftarrow top(pop^i)$
end for
return $best$.

Genetic algorithm

- Initial Population
  - for each residue, torsion angles (ϕ, ψ) are randomly chosen using APL
  - add a random small value $(\phi + \mathit{rand}(-1, 1), \psi + \mathit{rand}(-1, 1))$
- Crossover Operator
  - P1 from group A, P2 from group B + C
  - for each amino acid we use the information from P1 or P2 with a probability of 70% and 30%, respectively
- Next population
  - class A is promoted
  - individuals from the Local Search are inserted
  - class C is randomly chosen using APL

Genetic algorithm

- Local Search
  - Applied to each offspring generated by the crossover operator
  - For each residue it perturbs torsion angles with a probability of 10%
  - If a residue was chosen we used a greedy strategy to visit the neighbourhood of the angles with a size of one degree ($\phi - 1 \leq \phi \leq \phi + 1$, $\psi - 1 \leq \psi \leq \psi + 1$)
  - It is applied on ϕ, ψ, and side-chain torsion angles
  - It is computationally expensive, since each new solution must be fully evaluated.

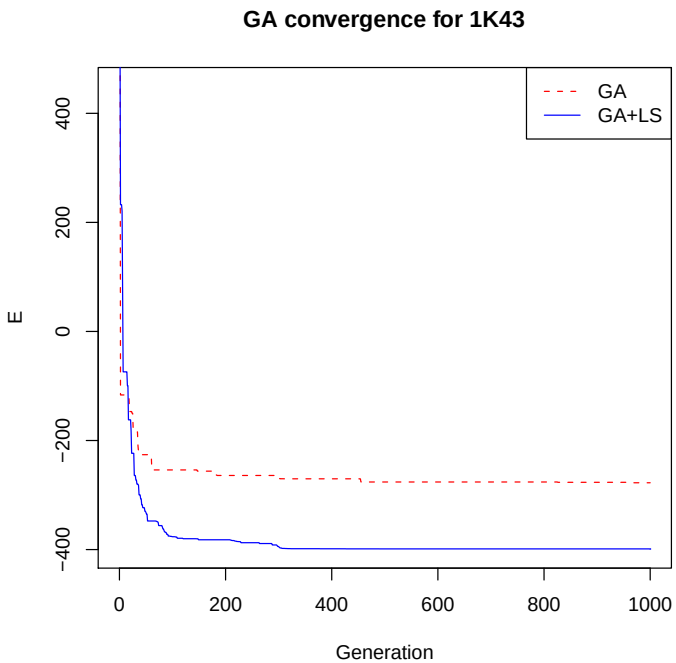
Computational Tests and Results Analysis

Computational Tests

- We use a set of six amino acid sequences taken from PDB:
  - 2EVQ(12 residues); 1K43 (14 residues); 1RPV (17 residues); 1L2Y (20 residues); 1DEP (15 residues) and 1ACW (29 residues)
- The energy function used is the Amber potential energy function
- The GA was ran on each sequence six times during two hours or 1,000 generations
- Analysis of the results was performed in terms of **structural analysis**, **secondary structure analysis** and **stereo-chemical analysis**.
- For this analysis we used the structures predicted with the lowest Potential Energy.

Computational Results

- The use of local search helps to improve the convergence of the algorithm in terms of quality and speed



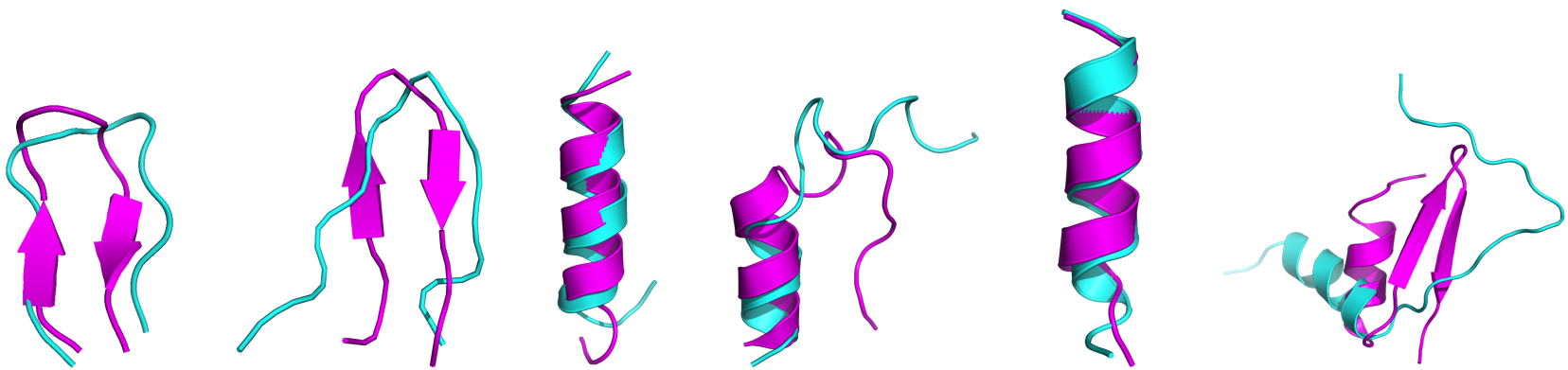
Protein	RMSD	E	$t(sec)$
2EVQ	3.29	-33.91	1672(76.9%)
1K43	4.15	-402.51	2823(79.7%)
1RPV	0.75	-719.26	7205(87.5%)
1L2Y	4.54	-211.2	5329(83.6%)
1DEP	0.85	-196.32	4406(84.4%)
1ACW	11.09	-138.41	7208(87.9%)

Computational Results

Structural analysis

Using the root mean square deviation (RMSD) compared with its native structure

- Individual helices and other secondary structures are well formed is most of the tested amino acid sequences
- β-sheets are not well formed, nevertheless these structures present small RMSD values



2EVQ 1K43 1RPV 1L2Y 1DEP 1ACW

Computational Results

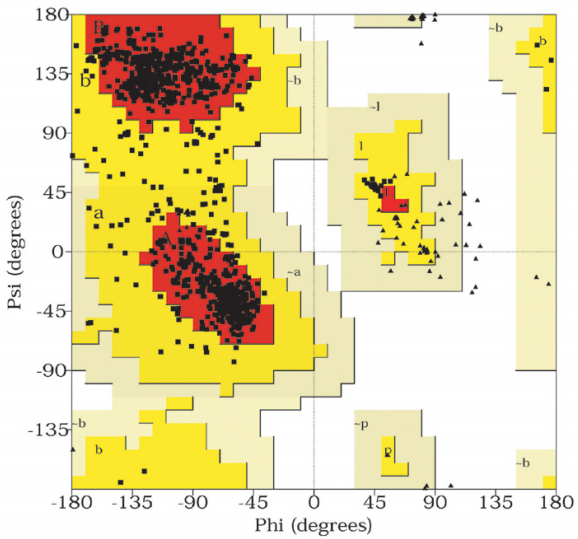
Secondary Structure and Stereo-chemical analysis

We analyse the patterns of hydrogen bonds compared with the native structure

	Secondary Structure Analysis				Ramachandran Plot Statistics			
Protein	β-sheet	α-helix	310 Helix	Other	(A)	(B)	(C)	(D)
2EVQ-E	50.0%	0.0%	0.0%	50.0%	87.5%	12.5%	0	0
2EVQ-P	0.0%	0.0%	0.0%	100.0%	100%	0	0	0
1K43-E	42.9%	0.0%	0.0%	57.1%	66.7%	33.3%	0	0
1K43-P	0.0%	0.0%	0.0%	100.0%	88.9%	11.1%	0	0
1RPV-E	0.0%	64.7%	0.0%	35.3%	86.7%	13.3%	0	0
1RPV-P	0.0%	64.7%	0.0%	35.3%	93.3%	6.7%	0	0
1L2Y-E	0.0%	35.0%	20.0%	45.0%	90.9%	9.1%	0	0
1L2Y-P	0.0%	45.0%	30.0%	25.0%	100%	0	0	0
1DEP-E	0.0%	80.0%	0.0%	20.0%	91.7%	8.3%	0	0
1DEP-P	0.0%	80.0%	0.0%	20.0%	91.7%	8.3%	0	0
1ACW-E	34.5%	24.1%	0.0%	41.4%	84%	16%	0	0
1ACW-P	0.0%	14.3%	0.0%	85.7%	96%	4%	0	0

Computational Results

- Stereo chemical analyses
  - The amino acid residues of the predicted 3D structures are mainly located in the most favourable region of the Ramachandran plot (red regions)
  - It suggests a small number of steric clashes
  - Sterio chemical properties of the predicted and experimental structures are comparable



Conclusion

Conclusion

- We have presented a Genetic Algorithm combined with a Local Search to the 3D PSP
- It is based on the use of knowledge about protein structures experimentally determined
- The Local Search helps the GA to escape from local minima and speed up the convergence to better solutions
- Predicted 3D structures are comparable to the experimental structure of the proteins
- Future Work
  - More advances techniques for Local Search are currently been explored
  - Use other Energy functions such as CHARMM and ECEPP
  - Refinement on the use of the information provided by the PDB
  - Further testing on larger protein sequences
  - Improve computational techniques

Acknowledgement

- Project Fondecyt 11121288, Chile
- Project CAPES 12216127, PROEX, Brazil
- FAPERGS 11/1410-1
- Organizing Committee of CEC2013 and IEEE