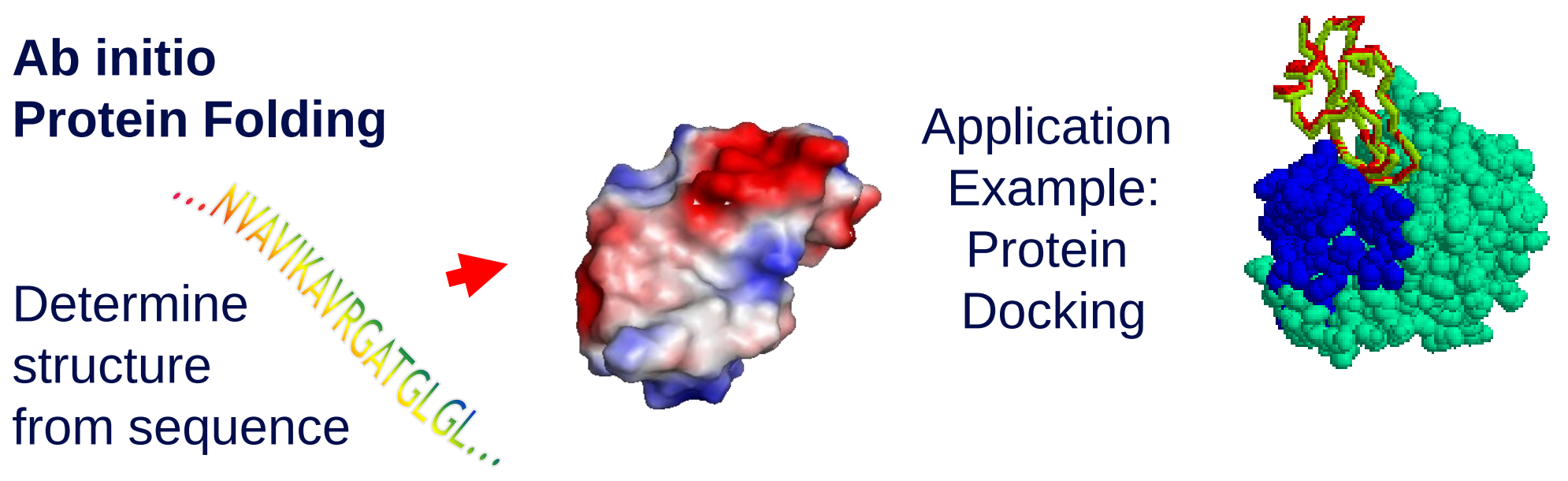
Ab Initio Protein Structure Prediction Using Conformational Search And Information From Known Protein Structures

[1,2] **Miguel M. F. Bugalho***, [1,2] **Arlindo L. Oliveira**

[1] **INESC-ID,**[2] **IST**

PORTUGAL

1 Abstract


Ab initio Protein Folding

Determine structure from sequence

Application Example: Protein Docking

The choice of the path to the near native conformation is a hard task. Our research is focused in two aspects:

- Fast generation of low energy conformations.
- Avoiding the creation of similar conformations.

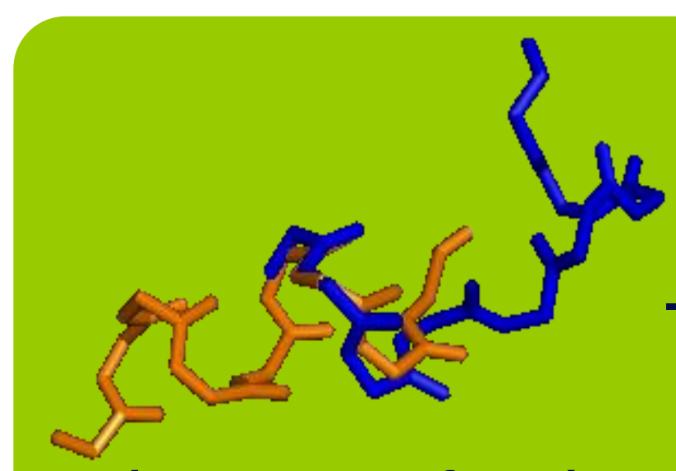
We will present a novel method that can efficiently generate low energy conformations. The proposed method uses the protein fragment library (9 AA) generated by ROSETTA[1]. We consider fragment overlap (3~4 AA). This reduces the number of degrees of freedom to only a fixed position and also enables the system to score the fragments using the degree of overlapping.

Motivation: If the fragments overlap, there is structural consistency between the two fragments that justify the usage of those fragments together. We use a statistical energy function, check for steric clashes[2]. All heavy atoms conformations with side chains placed using rotamer libraries[3].

Results and Conclusions: We can efficiently generate low energy conformations and, for smaller proteins, obtain near native conformations.

2 Algorithm

Basic algorithm



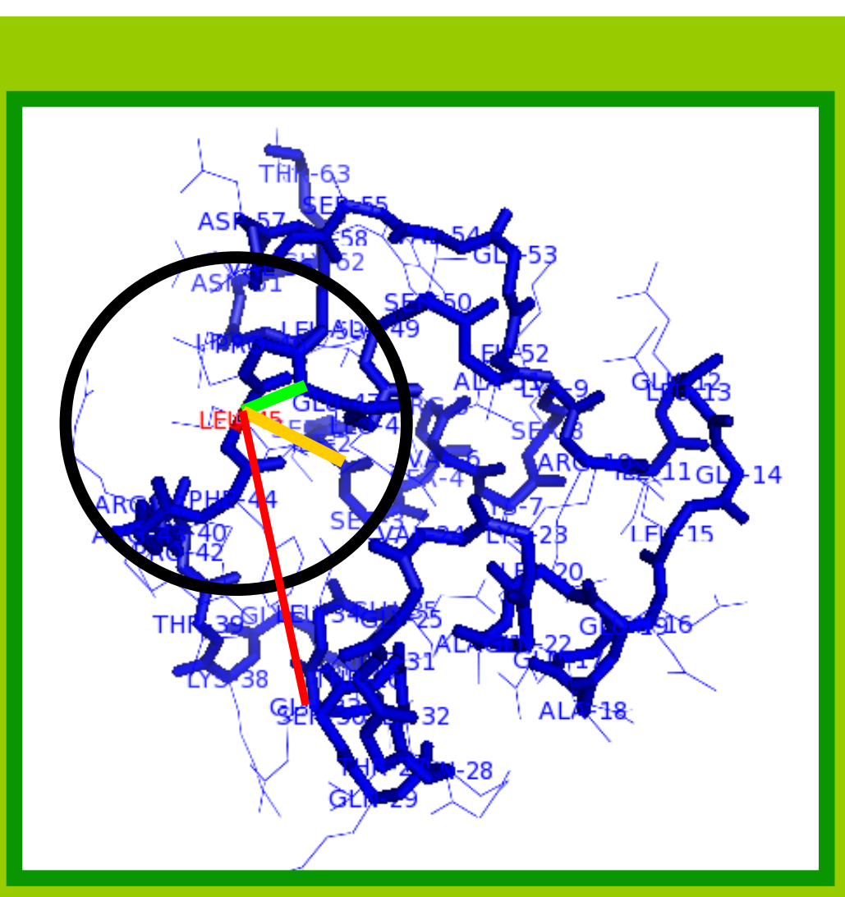
- Stochastic choice of fragments

-The score for the stochastic choice measures how well the fragment overlaps with the previous fragment

-Backtrack to previous fragment if a dead end is found

-After a conformation is found the algorithm backtracks to a previous state chosen stochastically from the search tree and constructs a new conformation

Scoring Function

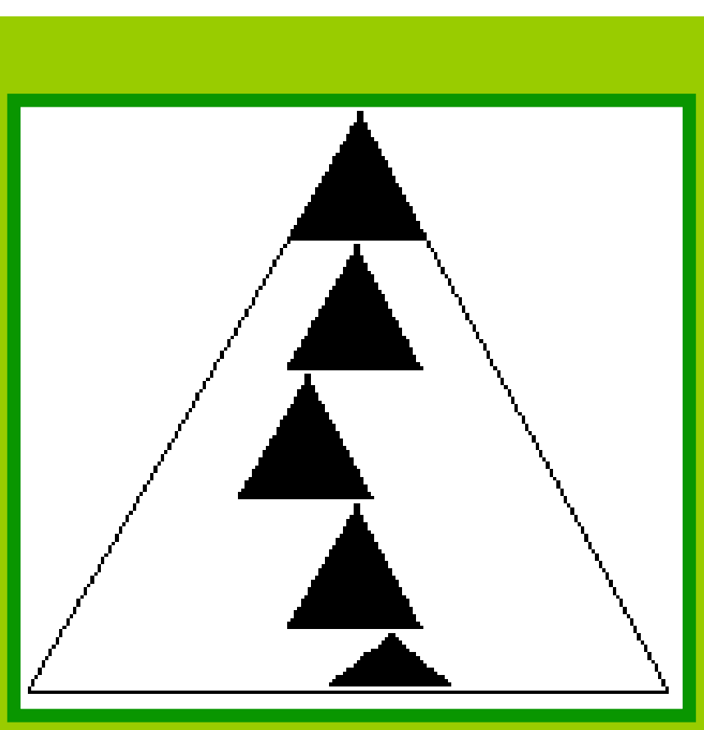


- Statistical scoring function to evaluate conformations
- Each fragment is rewarded according to its contribution
- Best conformations are chosen as base for new conformations

-Measures in the function (frequencies in proteins)

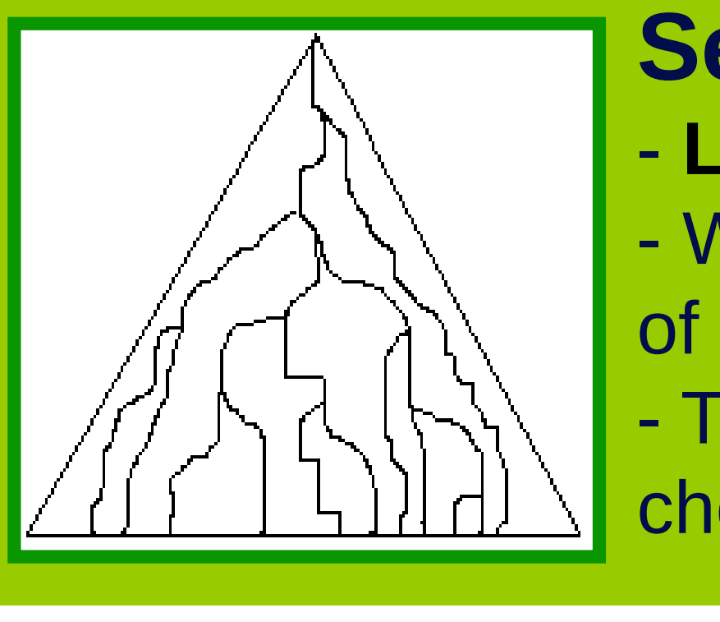
- **Buried State –** # of residues closer than cutoff (**Circle**)
- **Contacts –** Distance between AA (**Lines**) discrete slots
- **Radius of gyration –** Compactness of the conformation
- **Secondary structure –** Rewards fragments that present the secondary structure (PSIPRED[4])

Fragment Search (breadth-first/stochastic)



- Triangles represent a search in the available library fragments (generated using ROSETTA[1])
- Fragments are tested for clashe[2] and scored with current (fragment overlap) and previous information (scores in previous conformations)
- One fragment is chosen stochastically
- Backtrack starts if no fragments are available

Search tree (dead ends aren't represented)



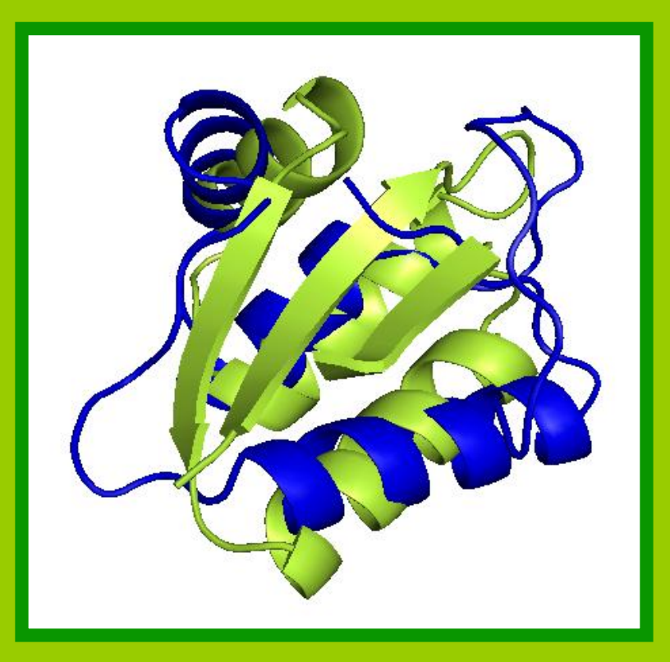
- **Lines** represent conformations, **points** along the line AA
- When a conformation is found the algorithm chooses one of the previous conformations as base for a new search
- The algorithm backtracks stochastically to a fragment choice (**forks**). Worse fragments have higher probability

3 Results and Conclusions

Results (all Ca atoms RMSD)

2000 Conformations	1ctf	1r69	3icb	1mol	1rro
Size	69	63	75	94	108
Type	α β	α	α	α β	α
Best	5.45	2.63	5.41	10.30	7.89
Best Pos	1450	28	347	273	1995
Top	6.17	6.72	6.62	11.63	12.84
Top 10 Best	5.72	5.29	6.58	11.63	11.63
10 Mean	6.16	6.23	6.60	12.54	13.56
Top 100 Best	5.66	2.63	6.07	11.63	11.62
100 Mean	6.54	6.20	6.93	12.79	12.24
Mean	8.17	6.67	7.76	13.60	12.18

1ctf

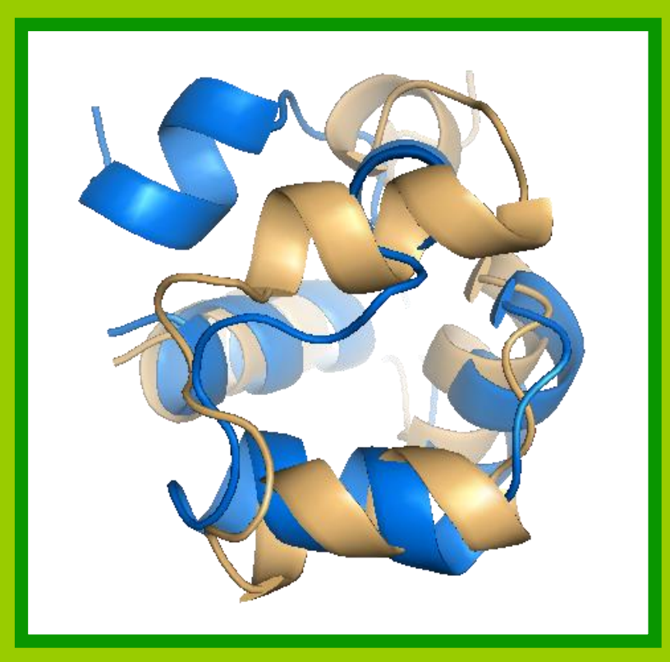


Best RMSD 5.454
Score 0.6835

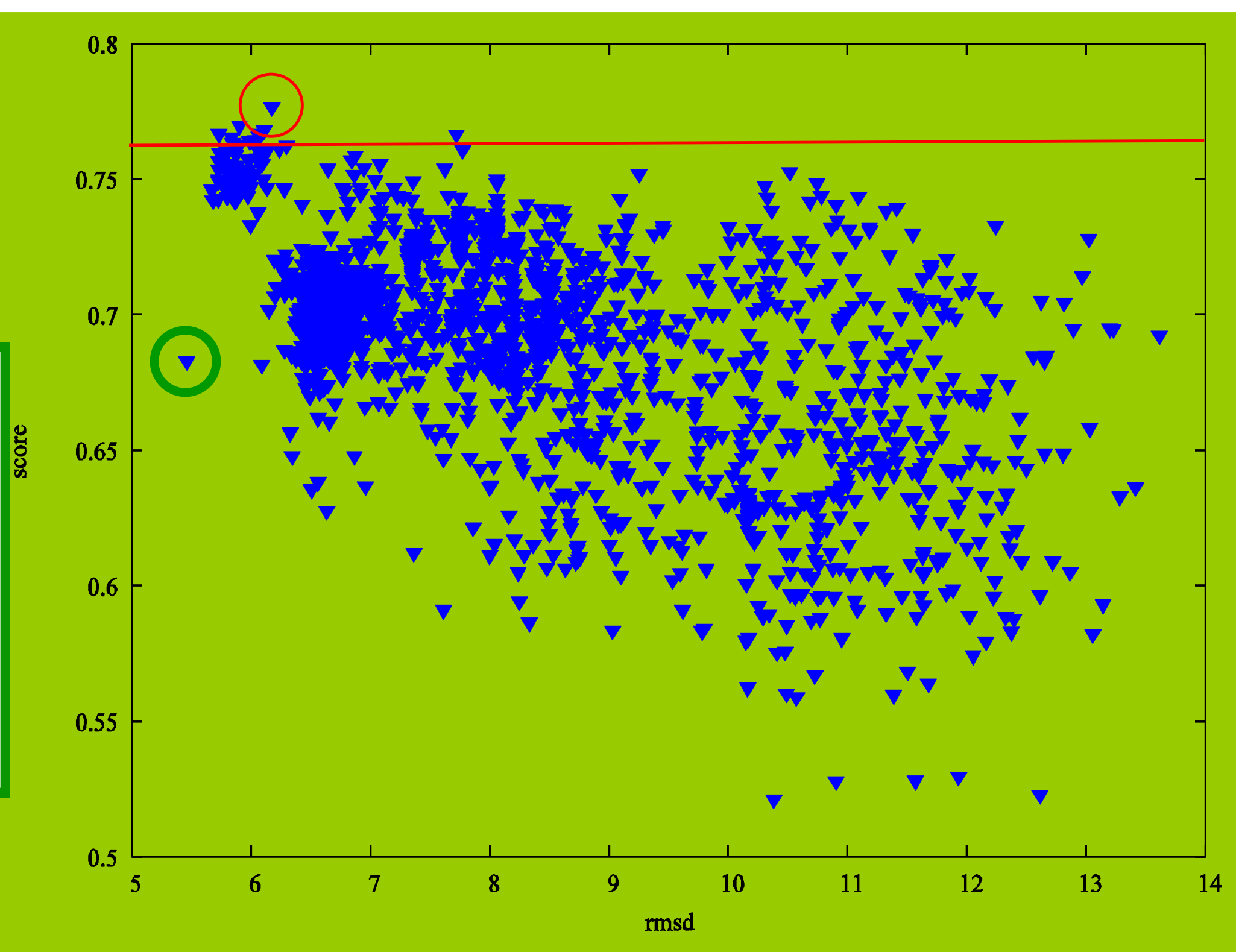
Top RMSD 6.167
Score 0.7771

1r69

Best RMSD 2.632
Score 0.7386



Top RMSD 6.724
Score 0.7613



Score vs RMSD
2000 1ctf decoys

- Best scored decoy in red
- Top 10 scored decoys above red line
- Best decoy generated in green (best RMSD)

Although the best decoy can't be differentiated using only score, a good decoy is normally scored highly.

Conclusions

- Conformations close to native fold can be found for small proteins
  - β strands and β sheet formation is hard to model
- The representations are physically correct, which facilitates refinement
- Efficient techniques are needed for finding the best generated conformation

Future Work

- Test different methods to create the fragment library (ex: variable size fragments)
- Improve generated conformation selection (ex: clustering) and use refinement
- Grid parallelization of the algorithm

4 References

[1] KT Simons et al. **Assembly of protein tertiary structures from fragments…** JMB 268:209-25,97
[2] M Bugalho, AL Oliveira **An efficient clash detection method…** BSB 08 (accepted) LNBI 5167
[3] RL Dunbrack Jr, M Karplus **Backbone-dependent rotamer library…** JMB 230(2):543–74,93
[4] DT Jones **Protein secondary structure prediction…** JMB 292: 195-202,99

Acknowledgements

This work was supported by INESC-ID and by the Portuguese Science and Technology Foundation by grant SFRH/BD/13215/2003 and project Biogrid POSI/SRI/47778/2002