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

Miguel M. F. Bugalho*, Arlindo L. Oliveira

INESC-ID, IST
PORTUGAL

Abstract
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.

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 – a 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 clash[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

Results and Conclusions

2000 Conformations 1ctf 1r69 3icb 1mol 1rro

<table>
<thead>
<tr>
<th></th>
<th>Size</th>
<th>Type</th>
<th>α</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>69</td>
<td>63</td>
<td>75</td>
<td>94</td>
</tr>
<tr>
<td>Best Pos</td>
<td>1450</td>
<td>28</td>
<td>347</td>
<td>273</td>
</tr>
<tr>
<td>Top</td>
<td>6.17</td>
<td>6.72</td>
<td>6.62</td>
<td>11.63</td>
</tr>
<tr>
<td>Top 10 Best</td>
<td>5.72</td>
<td>5.29</td>
<td>6.58</td>
<td>11.63</td>
</tr>
<tr>
<td>10 Mean</td>
<td>6.16</td>
<td>6.23</td>
<td>6.60</td>
<td>12.54</td>
</tr>
<tr>
<td>Top 100 Best</td>
<td>5.66</td>
<td>2.63</td>
<td>6.07</td>
<td>11.63</td>
</tr>
<tr>
<td>100 Mean</td>
<td>6.54</td>
<td>6.20</td>
<td>6.93</td>
<td>12.79</td>
</tr>
<tr>
<td>Mean</td>
<td>8.17</td>
<td>6.67</td>
<td>7.76</td>
<td>13.60</td>
</tr>
</tbody>
</table>

|          | 5.45 | 2.63 | 5.41 | 10.30 |
| Score vs RMSD 2000 1ctf decoys |
| Best RMSD 5.454 Score 0.6835 |
| Top RMSD 6.167 Score 0.7771 |

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

References

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