Large Scale Protein Side-Chain Packing Based on Maximum Edge-Weight Clique Finding Algorithms

Dukka Bahadur K. C.\textsuperscript{1} J. B. Brown\textsuperscript{1}
dukka@kuicr.kyoto-u.ac.jp jbbrown@kuicr.kyoto-u.ac.jp

Etsuji Tomita\textsuperscript{2} Jun’ichi Suzuki\textsuperscript{2} Tatsuya Akutsu\textsuperscript{1}
tomita@ice.uec.ac.jp jsuzuki@ice.uec.ac.jp takutsu@kuicr.kyoto-u.ac.jp

\textsuperscript{1} Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho
Uji, Kyoto 611-0011, Japan
\textsuperscript{2} Graduate School of Electro-communications, The University of Electro-
communications, Tokyo 182-85865, Japan

Keywords: protein side-chain packing, protein structure prediction, clique, clique finding algorithms

1 Introduction

The protein side-chain packing problem is computationally known to be NP-complete [1]. A number of approaches has been proposed for side-chain packing. As the size of the protein becomes larger, the sampling space increases exponentially. Hence, large scale protein side-chain packing

In this regard, we had also presented a maximum edge-weight clique based algorithm for protein side-chain packing which is able to compute the side-chain packing of a 500 residues long protein efficiently. However, the algorithm fails to produce a feasible solution for larger proteins because of
the nature of the search space.

We propose a new heuristic technique, for the side-chain packing algorithm based on maximum
edge-weight clique finding approach, that enables us to compute the side-chain packing of much larger proteins.

2 Method and Results

One of the approaches to overcome the search space intractability is to decompose the search space into smaller search space. Hence, we first divide the protein into a number of polymers, obtain edge weight subgraphs for each polymer and then calculate the cliques in each sub-graph.

2.1 Decomposition of Protein into Polymers

The original protein is first divided into a number of polymers. Although, deep insight into the problem may lead us to divide the protein at more plausible residues, in this work we divided the protein into \(n\) number of polymers with equal residues.

2.2 Graph Generation for each Polymer

The sampling of side-chain search space is done as in SPWCQ [2]. After the sampling step, both local consistency (the consistency of the vertices and edges of the polymer with vertices and edges of the same polymer) and global consistency (the consistency of the vertices and edges of the polymer with vertices and edges of other polymers) is maintained while generating graph.
In this way a maximum edge-weight graph corresponding to polymer \( P_i \) is generated. Once this process is repeated for all polymers from \( P_1 \) through \( P_p \), edge-weight subgraphs for all polymers are obtained. Although there is no theoretical proof that each polymer will have a maximum clique with the number of vertices being equal to the number of amino acids in each corresponding polymer, in all our computational experiments we were able to obtain a clique with the size of the number of amino acids in the polymer. In this way, subgraphs are generated for each polymer.

2.3 Clique finding

The maximum clique finding algorithm WCQ [4] developed by coauthor (Suzuki and Tomita) is applied for finding the maximum edge-weight cliques of the sub-graphs generated in the above step. Moreover, the final solution is obtained by taking the union of the vertices of the maximum edge-weight cliques of each sub-graph.

3 Computational Experiments

We present the results of computational experiments performed in order to assess the prediction accuracy of the new approach in Table 1.

<table>
<thead>
<tr>
<th>PDB</th>
<th>#Residue</th>
<th>( p=2 )</th>
<th>( p=3 )</th>
<th>( p=4 )</th>
<th>( p=5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1xwl</td>
<td>580</td>
<td>1.66</td>
<td>1.56</td>
<td>1.60</td>
<td>1.74</td>
</tr>
<tr>
<td>1gof</td>
<td>639</td>
<td>1.63</td>
<td>1.46</td>
<td>1.69</td>
<td>1.70</td>
</tr>
<tr>
<td>1aa6</td>
<td>696</td>
<td>1.58</td>
<td>1.68</td>
<td>1.61</td>
<td>1.7</td>
</tr>
<tr>
<td>1a8i</td>
<td>812</td>
<td>1.56</td>
<td>1.63</td>
<td>1.63</td>
<td>1.75</td>
</tr>
<tr>
<td>1fyy</td>
<td>874</td>
<td>1.42</td>
<td>1.53</td>
<td>1.63</td>
<td>1.54</td>
</tr>
</tbody>
</table>

4 Conclusion and Discussion

We have developed a new approach for large scale side-chain prediction based on a heuristic technique to divide the original graph into subgraphs. One of the future directions of the research is to incorporate more bio-chemical knowledge in order to break the protein at relevant residues.

References


