Selection of Proper Template Complex for Modeling Protein-Protein Complex

Naoshi Fukuhara\textsuperscript{1} Takeshi Kawabata\textsuperscript{1}
\texttt{n-fukua@is.naist.jp} \texttt{takawa@is.naist.jp}
Nobuhiro Go\textsuperscript{1,2}
\texttt{ng@is.naist.jp}

\textsuperscript{1} Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan
\textsuperscript{2} Centre for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, 8-1 Umemidai, Kizu, Souraku, Kyoto, 619-0215, Japan

Keywords: protein-protein interaction sites, three-dimensional structure, homology

1 Introduction

There is a gap in knowledge between complexes of known three-dimensional structure and those known from other experimental methods such as affinity purifications or yeast two hybrid. In order to bridge this gap, it is important to extrapolate interaction information from one complex structure to homologues of the interacting proteins.

Two approaches have been developed recently to predict protein-protein interaction by homology modeling and threading techniques [1, 2]. However, Aloy \textit{et al.} performed the statistics of known protein complexes, and concluded that if sequence identities are more than 30-40\%, interactions are almost invariable, otherwise interactions can be different [3]. Their result suggests selection of proper template complex structure is important for modeling complex structures. In this study, we also predict protein-protein interaction sites using homology modeling, and try to find effective features for filtering out the improper template structures.

2 Data and Methods

2.1 Preparation of dataset

A library dataset of interacting chain pairs is obtained from the complex structures registered in the PQS server [4]. It includes only complexes annotated as “biological oligomer”. From the library dataset, a representative test set are extracted with 40\% sequence identities or less. Two types of test sets are prepared: 291 hetero-dimers and 1562 homo-dimers (sequence identity between interacting chains is less than 50\% and more than 95\%, respectively).

2.2 Prediction Algorithm

For each two chains from query dimer, BLAST search is run against the library. If the obtained homologous two chains also form complex in the same PQS file, this complex is adopted as a candidate of template structure. Next, the interaction sites are calculated in the template structure on the criteria where atomic distance is smaller than 4Å. The corresponding query sites for the interaction template sites in the alignment are predicted as the interaction sites. The performance of the prediction is assessed by correlation coefficient (CC) between the predicted and true interaction sites.
3 Results and Discussion

Figure 1: The relationship between sequence identity of complexes and correlation coefficient of the prediction. Query is homo-dimer (left). Query is hetero-dimer (right). Each point represents a pair of query and template.

The relationships between sequence identity and prediction accuracy are summarized in Figure 1. In this prediction, the maximum (SeqID ≤ 95%) and minimum (E-value ≥ 0.0001) threshold was set. As the SeqID becomes lower than about 40%, the number of the templates which falsely predict interaction sites rapidly increase. Comparing the case of the homo-dimers (left) with that of hetero-dimers (right), it is clear that the former have the enormous improper templates than the latter. We observed that most of the reasons for the improper templates are artificial crystal packing. These results are consistent with Aloy et al. [3].

Several features are tried for filtering out the improper templates for the prediction of hetero-dimers. We found that alignments of improper cases are often too short to include all of the interaction sites of template. Figure 2 shows the frequency distribution of aligned interaction sites of template structure for proper templates (CC ≥ 0.2) and improper templates (CC < 0.2). It shows this feature can effectively discriminate improper templates. The origin of short alignment is not only weak sequence similarity, but also domain insertion or deletion, or cleavage by protease. And we found a hetero-dimer medeled by a homo-multimer template often leads to a poor result. We are now planning to modify an algorithm for solving these problems.

References