FUJI Database Provides Functional and Structural Annotation of Proteins On Demand

Keiichi Homma¹
khomma@lab.nig.ac.jp

Hideaki Sugawara¹
hsugawar@genes.nig.ac.jp

¹ Targeted Proteins Research Program, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Keywords: functional prediction, annotation, database, protein structure

1 Introduction

With the number of protein sequences exponentially growing, it is becoming increasingly hard to add relevant information to all the existing sequences and make it publicly available in a timely fashion. The situation is especially difficult for functional annotation of proteins, as the labor-intensive manual curation process can hardly handle the deluge of sequences. An on-demand service to give information including functional annotation is superior to the traditional database approach if it can offer reliable information using updated data as rapidly as stored data are retrieved. The NCBI site allows the user to conduct a BLAST search of a specified protein sequence against MMDB-validated PDB sequences [1], but only presents a three-dimensional view of PDB structures themselves, with no information on unaligned regions or other sequence features. On the other hand, the Protein Annotation Workflow [2] supplies a wide range of structural information on demand, but it takes more than 20 min to reply and does not give a functional prediction. As part of the Targeted Proteins Research Program administered by the MEXT, Japan, we aimed to build a system that expeditiously provides accurate structural information and a functional prediction. Currently the system returns results of various analyses in response to any protein sequence in a few minutes.

2 Method and Results

We have built a system and made it publicly available as the Functionally Annotated Japanese Protein Structural Information (FUJI) database (http://fujidb.genes.nig.ac.jp/fujidb/). The FUJI database first requests the user to specify a protein sequence (Fig. 1). The system then conducts a number of analyses including alignments to structural domains and sequence motifs, prediction of disordered regions and subcellular localization, and functional annotation with the use of the structural information. In a path distinct from the fully automated process, the system selects a candidate of the most conserved region derived by homolog alignments and asks the user to modify it. The information added is displayed linearly, with the N-terminus of the protein shown at the left. If structural alignments exist, the aligned regions can be displayed three dimensionally.

Specifically, the FUJI database carries out alignments to SCOP, Pfam, SMART domains and PDB structures. The system also makes predictions of disordered, SEG, and coiled-coil regions, a signal sequence, trans-membrane domains (TMDs), subcellular localization, and protein function. In a separate data flow, FUJI selects homologs by BLAST, aligns them, identifies the most conserved region, and prompts the user to modify the set of homologs and to manually alter the conserved region itself.

The first screen display shows the garnered information linearly. Clicking on the 3D Display button, if any, the user can examine the aligned structure in a new viewer with a suitable angle and magnification. Unaligned regions, gaps in PDB coordinates, the inside and outside of the protein molecule (as judged by solvent accessibility), and the conserved region are also shown.
3 Discussions

The FUJI database is unique in the rapidity of the on-demand service it provides. From the database, the user can swiftly get a wealth of structure-related information on any protein whose sequence is known. Furthermore, the database makes a functional prediction, utilizing structural information. The accuracy of the automatic prediction system was shown to be higher than that obtained by the BLAST best-hit method. Making use of the information, the user can give an intelligent guess about functions, functionally important regions, and possible sites to introducing mutations, among many others.

We aim to keep the database up-to-date so that the user can get the latest information. We are also making efforts to improve the functional prediction system through the development of an automatic pseudogene identification algorithm and a system optimized for each kingdom of life.

References