ABSTRACT: With the advent of the information age, bioinformatics has been rapid development in many fields to be fully applied. This paper describes the establishment and development of common database of protein, protein structure analysis, protein secondary and tertiary structure prediction; The application of bioinformatics in protein structure analysis is comprehensively summarized, which provides theoretical basis for the wide application of bioinformatics in protein engineering.

1. INTRODUCTION

Bioinformatics provides a shortcut for the protein function prediction and structural analysis by directly using gene or protein sequences. Which is formed by the interaction of biological and computer science and applied mathematics. It is processed by the data acquisition, processing, storage, retrieval and analysis, to explain the biological significance of the data. As the genome and proteome research provides a very rich data, we need to make a high degree of automation of these data management, establish a strict database and write the corresponding software. This work greatly promoted the development of bioinformatics, and bioinformatics will also play a special role in the research of structure prediction (Hagen, 2000).

The emergence of bioinformatics has made the protein project into a new era. In the future, more detailed knowledge of protein structure and function, as well as advancements in high throughput technology, may greatly expand the capabilities of protein structure prediction.

Protein structure prediction is an important task in bioinformatics. To a large extent, the biological function of protein depends on its spatial structure, so it is very important to predict the structure of proteins. Sometimes a possible new gene cannot be found by searching for any homologous sequence, namely orphan gene, and for proteins encoded by those genes, bioinformatics technology can be used to search for homologous genes based on the homologous comparison of structure or predict its advanced structure to infer its function directly.

2. PROTEIN STRUCTURE DATABASE

According to the data in the database based on the analysis of the structure and function of the natural protein, the spatial structure and biological function of a certain amino acid sequence can be predicted, and the amino acid sequence and spatial structure of the protein can be designed according to the specific biological function.

The basic three-dimensional structure of the protein database is Protein Data Bank (PDB; http://www.pdb.org/), founded by Brookhaven National Laboratory in New York in 1971, which is the database of the most important biological macromolecules (proteins, nucleic acids and sugars) three-dimensional structure (Sussman et al., 1998; Berman et al., 2000; Westbrook et al., 2003). The database is accurate coordinate data collected in the protein structure by X-ray diffraction and nuclear magnetic resonance (NMR) experiment measured. Protein structure data in PDB are widely used in studies of protein function and evolution, and they serve as a basis for protein structure prediction (He et al., 2014).

PDB Finder database (http://www.cmbi.kun.nl/swift/pdbfinder/) is built on the basis of PDB, DSSP, HSSP, the two level library, which contains the PDB sequence, the author, R factor, resolution, Secondary Structure, etc. these information is not easy to read directly from the PDB, with the PDB library every time a new version, PDBFinder automatically generated in EBI.

Homology Derived Secondary Structure of Proteins (HSSP; http://www.sander.embl-heidelberg.de/hssp/) is the secondary structure database based on the homology of the two protein structure. Each PDB project has a corresponding HSSP file (Dodge et al., 1998). The database also provides the homology of all protein sequences in the SWISS-PROT database.
The Structural Classification of Proteins database (SCOP; http://scop.mrc-lmb.cam.ac.uk/scop/) is a comprehensive ordering of all proteins of known structure, according to their evolutionary and structural relationships, which is obtained by manual classification, in which the proteins of known structure are hierarchical classified (Murzin et al., 1995; Andreeva et al., 2004). This resource allows users to analyze whether the query protein and known structural protein have similarities. Protein domains in SCOP are hierarchically classified into families, superfamilies, folds and classes.

Molecular Modeling Database (MMDB; http://www.ncbi.nih.gov/Structure/MMDB/mmdb.shtml), This is a three-dimensional structure database used by the Entrez search tool (Wang et al., 2000). ASN1 format, which reflects the structure and sequence data of PDB Library. Meanwhile, a complete set of 3D structure display program Cn3D is provided by NCBI.

MUFOLD-DB (http://mufold.org/mufolddb.php), a web-based database, to collect and process the weekly PDB files thereby providing users with non-redundant, cleaned and partially predicted structure data. For each of the non-redundant sequences, the SCOP domain classification and predict structures of missing regions are annotated by loop modelling. In addition, evolutionary information, secondary structure, disorder region, and processed three-dimensional structure are computed and visualized to help users better understand the protein (He et al., 2014).

3. PROTEIN STRUCTURE PREDICTION

Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its folding and its secondary, tertiary, and quaternary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes) (https://en.wikipedia.org/wiki/Protein_structure_prediction).

3.1 Protein secondary structure prediction

Protein secondary structure prediction is a set of techniques in bioinformatics that aim to predict the local secondary structures of proteins based only on knowledge of their amino acid sequence only. Which plays a vital role and acts as an intermediate in solving tertiary structures; which provides an insight into protein function (Kloczkowski et al., 2002). For proteins, a prediction consists of assigning regions of the amino acid sequence as likely alpha helices, beta strands (often noted as "extended" conformations), or turns. The different amino acid residues have different tendencies for the formation of different secondary structures. The success of a prediction is determined by comparing it to the results of the DSSP algorithm (or similar e.g. STRIDE) applied to the crystal structure of the protein. Specialized algorithms have been developed for the detection of specific well-defined patterns such as transmembrane helices and coiled coils in proteins (Mount, 2004). The prediction accuracy of a single sequence of secondary structure is about 60% ~80% (Petersen et al., 2000). The best modern methods of secondary structure prediction in proteins reach about 80% accuracy (Pirovano & Heringa, 2010); this high accuracy allows the use of the predictions as feature improving fold recognition and ab initio protein structure prediction, classification of structural motifs, and refinement of sequence alignments. 10 kinds of important and effective methods were tested and analyzed using the unified standard, the comparison results show: the prediction effect of APSSP2, SsPro2.0 and PSIPRED method is better, the accuracy of the prediction is more than 80% (Zhang et al., 2003).

So far, more than 20 different secondary structure prediction methods have been developed. First algorithms was Chou-Fasman method, which relies predominantly on probability parameters determined from relative frequencies of each amino acid's appearance in each type of secondary structure (Chou et al., 1974). The next notable program was the GOR method, is an information theory-based method. Which employs the more powerful probabilistic technique of Bayesian inference (Garnier et al., 1978). Another big step forward, was using machine learning methods. First artificial neural networks methods were used. As a training sets they use solved structures to identify common sequence motifs associated with particular arrangements of secondary structures. These methods are over 70% accurate in their predictions, although beta strands are still often underpredicted due to the lack of three-dimensional structural information that would allow assessment of hydrogen bonding patterns that can promote formation of the extended conformation required for the presence of a complete beta sheet (Mount, 2004). PSIPRED and JPRED are some of the most known programs based on neural networks for protein secondary structure prediction (Cuff, 2000).

The predicted secondary structural states are not cross validated by any of the existing servers. Hence, information on the level of accuracy for every sequence is not reported by the existing servers. This was overcome by NNvPDB, which not only reported greater Q3 but also validates every prediction with the homologous PDB entries.
NNvPDB is based on the concept of Neural Network, with a new and different approach of training the network every time with five PDB structures that are similar to query sequence. The average accuracy for helix is 76%, beta sheet is 71% and overall (helix, sheet and coil) is 66% (Sakthivel et al.,2015).

3.2 Protein tertiary structure prediction

Protein tertiary structure prediction has been an important scientific problem for few decades, especially in bioinformatics and computational biology (Eisenhaber et al., 1995). Despite more and more native structures are included in protein data bank (PDB) database, the gap between the sequenced proteins and the native structures is still enlarging due to the exponential increase of protein sequences produced by large-scale genome and transcriptome sequencing. It is estimated that <1% of protein sequences have the native structures in PDB database (Rigden et al.,2009). Therefore, accurate computational methods for protein tertiary structure prediction that are much cheaper and faster than experimental structure determination techniques are needed to reduce this large sequence structure gap. Furthermore, computational structure prediction methods are important for obtaining the structures of membrane proteins whose structures are hard to be determined by experimental techniques such as X-ray crystallography (Yonath et al., 2011).

At present, the main methods of 3D spatial structure prediction are homology modeling, fold type recognition and ab initio prediction. Homology modeling is the most successful and useful protein structure prediction methods, which can be used to explain the experimental data, proceed mutants design and drug design. Today, SWISS-MODEL is one of the most widely used structure modelling web servers world-wide, with more than 0.9 million requests for protein models annually (Biasini et al.,2014). Several other groups have developed similar systems for automated homology modeling [e.g. ModPipe (http://www.salilab.org) (Sanchez et al., 1998), CPHmodels (http://www.cbs.dtu.dk/services/CPHmodels/) (Lund et al.,1997), 3D-JIG SAW (http://www.ncbi.nlm.nih.gov/3djigaw/) (Bates et al., 2001), ESYPred3D (http://www.fundp.ac.be/urbm/bioinfo/esypred/) (Lambert et al.,2002) or SDSC1 (http://c1.sdsc.edu/hm.html).

At the beginning of 1990s, the outstanding progress of protein structure prediction is the recognition method of protein folding type, also known as "threading" or "reverse folding" (Sippl et al.,1999). Its main contribution is to predict the structure of proteins that have no obvious homology but also have similar structure. Current applications are mainly on the sequence-based methods. Evolutionary information is obtained by comparison of sequences with known structures, such as PBLAST (Altschu et al., 1997) and hidden Markov model [HMM]-based methods (Karplus et al.,2001). The structure of spike protein (S2) in SARS virus and HIV-1 gp41 were compared, analysis reveals that, although the protein responsible for viral-induced membrane fusion of HIV-1 (gp41) differs in length, and has no sequence homology with S2. The two viral proteins share the sequence motifs that construct their active conformation (Kliger et al., 2003). The structural features contribute to further design viral suppression drugs.

The ab initio prediction method is not dependent on the existing structural data information, using molecular dynamics theory to predict and infer structural information directly from the protein sequence. At present, most ab initio prediction methods are used to search for spatial structure of proteins, such as MonteCarlo, simulated annealing, genetic algorithm, and so on. The energy of each structure is also evaluated, in order to find the lowest free energy. Jones presents predictions approach based on fragment, by using simulated annealing algorithm (SA), supe-secondary structural fragments were assembled to achieve protein de novo prediction (Jones et al.,2003). To address the searching problem of protein conformational space in ab-initio protein structure prediction, a novel method using abstract convex underestimation (ACUE) based on the framework of evolutionary algorithm was proposed. Computing such conformations, essential to associate structural and functional information with gene sequences, is challenging due to the high-dimensionality and rugged energy surface of the protein conformational space (Hao et al.,2015).

4. ASSESSMENT FOR PROTEIN STRUCTURE PREDICTION

Methods for protein structure prediction are flourishing and becoming widely available to both experimentalists and computational biologists. But, how good are they? What is their range of applicability and how can we know which method is better suited for the task at hand? Some evaluation methods focus on these questions, such as the world-wide Critical Assessment of Techniques for Protein Structure Prediction(CASP; http://www.predictioncenter.org/experiment(Moult et al.,1995; Tramontano et al.,2008)), automatic evaluation servers such as EV aluation of Automatic protein structure prediction(EVA; http://cubic.bioc.columbia.edu/eva/) (Koh et al.,2003) and Livebench(http://bioinfo.pl/meta/livebench.pl) (Fischer et al.,2000).
5. OUTLOOK

Biological information science is a new cross subject, is still continuous progress and improvement, it is not the "panacea" of study in molecular biology, it is impossible to completely replace the experimental operation. The analysis and prediction of bioinformatics is based on the knowledge of molecular biology. With the full and effective use of previous theoretical knowledge, which make reasonable inference. Therefore, there may be errors, which still need to be verified and added in the laboratory work.

REFERENCES


