

Structural Genomic View of Microbial Genomes

Taesung Moon¹, Jin Kak Lee^{1,2}, Won Seok Han¹, Han Su Choi¹, Min Su Han¹, Ki Dong Eom^{1,2},
Min Kyung Lee¹, and Chang No Yoon^{2*}

¹Nanormics Inc., ²Bioanalysis and Biotransformation Research Center, Korea Institute of Science and Technology

Abstract

At present, hundreds of microbial genomes have been sequenced, and hundreds more are currently in the sequencing pipeline. As the amount of genome data is expanding, researchers are much in need of tools that can process huge amount of sequence data. Here, we will discuss about several bioinformatics tools and their applications.

Introduction

The number of sequences in GenBank (1) which is maintained by NCBI (National Center for Biotechnology Information) is increasing twice every 14 months. However, the efforts on genome project may become useless if the huge amount of sequencing data cannot be accessed by researchers. Therefore, many organizations including NHGRI (National Human Genome Research Institute, 2) make a lot of efforts to produce computer programs and user interfaces in order that the researchers can access the database more easily. In this presentation, we introduce Genome Annotation System for microorganism based on 3D protein structure modeling and show some applications of these systems, for examples, protein design and engineering, and prediction of essential genes for new targets for antibacterial drugs.

Genome Annotation

The main goal of bioinformatics is to collect biological information and systematically integrate them. This is the reason why so many bioinformatics-related companies and organizations build the genome browsers.

Requirements for genome browser :

- Genome browser should contain large quantity of biological information.
- Contents of Genome browser should have a clear definition.

- Collected information should be organized systematically.
- Convenience for the users should be guaranteed.

Nanormics Inc. (www.nanormics.com, 3), one of the bioinformatics companies, has the proprietary genome browser which meets the requirements, and provides a commercial version of this genome browser, NanoRNGAS.

Regulatory Network based Genome Annotation System (RNGAS)

RNGAS consists of three topics; structural genomics, functional genomics and comparative genomics. Given a full genome sequence, protein coding regions are identified and their putative functions for those regions are, then, defined based on a series of conventional annotation procedures and then, predicting uncharacterized operons and building a global regulatory network (Fig.1) are performed. The network for a given genome includes clusters for primary and secondary metabolite encoding genes, regulon, modulon and stimulon as well as protein-protein interactions (4,5).

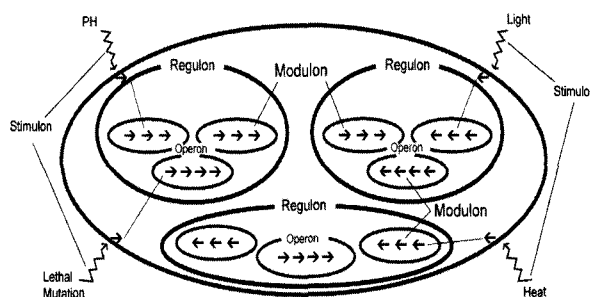


Fig. 1. Concept of Global Regulatory Network in NanoRNGAS

Protein Structure Modeling

The functional characteristic of protein sequences is central to many problems in biology. This task is facilitated by an accurate 3D structure of the protein of interest. The gap between the numbers of protein sequences and structures is increasing rapidly, accelerated by the completion of numerous genome projects now flooding into public databases. To fill this gap, the model structure database (NanoModelDB, 6) is generated by the NanoFold engine. High-throughput, automatic protein modeling should considerably increase our access to protein structures other than those determined by experimental techniques. NanoModelDB is updated monthly to ensure that new data can be added to the protein

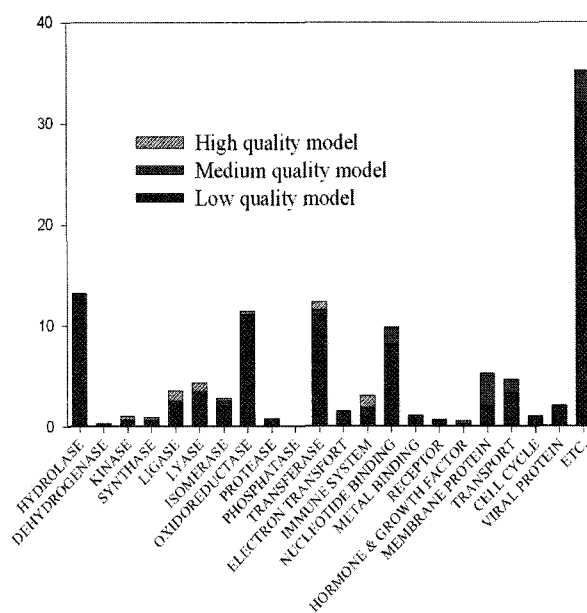


Fig. 2. Functional distribution of protein models in NanoModelDB

model structure database. We have selected 289 species and carried out automated modeling. Total of 594,201 (coverage: 55.38%) model structures are available now.

Protein Design and Engineering

The protein structures are essential for understanding of numerous biological processes from enzymatic reactions to immune systems. Nanormics Inc. has developed techniques for high-throughput screening system based on quantitative protein-ligand interaction analysis. These technologies have been applied to NanoDesign, the structure-based protein design and engineering systems. These systems enable the production of new antibodies as well as the modification of existing proteins for the purpose of producing new ones that exhibits new substrate specificity, increased thermal, pH, and solvent stabilities, from the microbial resources.

Prediction of Essential Gene for Drug-Target

Essential genes are genes that are indispensable to support cellular life. The proteins expressed by essential genes carry out basic functions of metabolism or involve in cell survival such as cell protection from external attacks. Essential genes not only give a clue for many biological questions but also are used for various purposes, for example, target for the development of new antibiotics. Many experimental and bioinformatics methods have been developed for the identification of essential genes. NanoEssence, predictor of essential genes, is developed on the basis of NanoRNGAS, NanoModelDB, and NanoDesign.

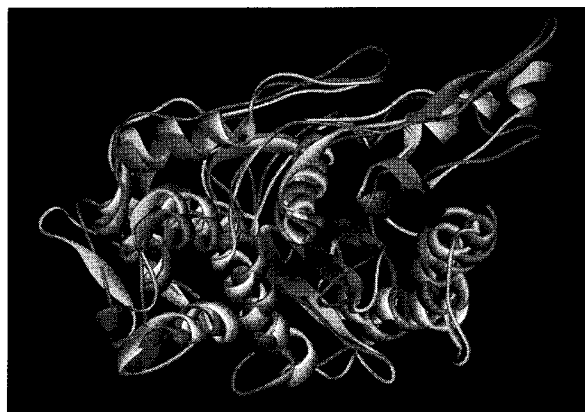


Fig. 3. PBP1 (penicillin binding protein 1: light blue) is not essential gene because the function of PBP2a (methicillin resistance protein: grey) is same as that of PBP1.

Conclusion

Genome annotation and protein modeling systems are very useful for analyzing and understanding of microbial genome data. And protein design and engineering systems enable us to produce new proteins from the microbial resources, which can be applied to industrial, medicinal, and agricultural fields.

References

1. GenBank, National Center for Biotechnology Information <http://www.ncbi.nlm.nih.gov>
2. National Human Genome Research Institute (NHGRI) <http://www.genome.gov>

3. Nanormics Inc. <http://www.nanormics.com>
4. DIP, Database of Interacting Proteins <http://dip.doe-mbi.ucla.edu>
5. BIND, Biomolecular Interaction Network Database <http://www.bind.ca>
6. NanoModel Database System. <http://www.nanormics.com>