Volume 2, Issue 3

18th July 2011

TezuBioin

(A Bioinformatics and Biotechnology e-Newsletter)

Bioinformatics Infrastructure Facility
Department of Molecular Biology and Biotechnology
School of Science and Technology
Tezpur University Newsletter

Inside this issue:

Upcoming Events:

Free Webinar on "60 minutes Minitab training on how to run a 1-Sample t-Test" on Jul 28, 2011 2:00 PM EDT (GMT -4:00) and Aug 9, 2011 11:00 AM EDT (GMT -4:00).
For registration:
http://www.minitab.com/en-US/training/web-events/

Cover Story: Bioinformatics and its use

Bioinformatics is the application of computer science and information technology to the field of biology.

The term bioinformatics was coined by Paulien Hogeweg and Ben Hesper in 1978 for the study of informatic processes in biotic systems. Its primary use since at least the late 1980s has been in genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing.

The primary goal of bioinformatics is to increase the understanding of biological processes. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies and the modeling of evolution.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data.
Over the past few decades rapid developments in genomic and molecular technologies and developments in information science have combined to produce a tremendous amount of information related to molecular biology. It is the name given to these mathematical and computing approaches used in-depth understanding of biological processes.

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different nucleotide and protein sequences to compare and create, gene finding, drug designing and viewing 3-D models of protein structures.

Special Interest: What is a Biological Database?

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technologies, and computational analyses. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.
Relational database concepts of computer science and Information retrieval of digital libraries are important for understanding the biological databases. Biological database design, development, and long-term management is a core area of the discipline of bioinformatics. Data contents include gene sequences, textual descriptions, attributes and ontology classifications, citations, and tabular data. These are often described as semi-structured data, and can be represented as tables, key delimited records, and XML structures. Cross-references among databases are common, using database accession numbers.

Biological databases are important tool in assisting scientists to understand and explain a host of biological phenomena from the structure of bio-molecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps facilitate the fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in the history of life.
Biological knowledge is distributed amongst many different general and specialized databases. This sometimes makes it difficult to ensure the consistency of information. Biological databases cross-reference other databases with accession numbers as one way of linking their related knowledge together.

Cheminformatics: Lipinski rule of 5

Lipinski's Rule of Five is a rule of thumb to evaluate drug-likeness, or determine if a chemical compound with a certain pharmacological or biological activity has properties that would make it a likely orally active drug in humans beings. The rule was formulated by Christopher A. Lipinski in 1997 based on the observation that most medication drugs are relatively small and lipophilic molecules.
The rule describes molecular properties important for a drug's pharmacokinetics in the human body, including their absorption, distribution, metabolism, and excretion ("ADME"). However, the rule does not predict if a compound is pharmacologically active.
The rule is important for drug development where a pharmacologically active lead structure is optimized step-wise for the increased activity and selectivity, as well as drug-like properties as described by Lipinski's rule.

Lipinski's rule says that, in general, an orally active drug has no more than one violation of the following criteria:

- Not more than 5 hydrogen bond donors (nitrogen or oxygen atoms with one or more hydrogen atoms)
- Not more than 10 hydrogen bond acceptors (nitrogen or oxygen atoms)
- A molecular mass not greater than 500 daltons
- An octanol-water partition coefficient log P not greater than 5
- Rotatable bonds less than 10

All numbers are multiples of five, which is the origin of the rule's name.
Besides, there was improvements in the Lipinski's Rule of Five to evaluate druglikeness better, the rules have spawned many extensions which are:

- Partition coefficient log P in -0.4 to +5.6 range
- Molar refractivity from 40 to 130
- Molecular weight from 160 to 500
- Number of atoms from 20 to 70 (includes H-bond donors [e.g.;OHs and NHs] and H-bond acceptors [e.g.; Ns and Os])
- Polar surface area no greater than 140 Å^2

Over the past decade Lipinski's profiling tool for druglikeness has led to further investigations by scientists to extend profiling tools to lead-like properties of compounds in the hope that a better starting point in early discovery can save time and cost. Some compound searching sites, such as the extensive searching tools on the ASDI website, now allow using Rule-of-5 and other properties to rapidly identify compounds that may be more desirable for high throughput screening and for parallel synthesis efforts.

Biological Databases: DNA Databank of Japan (DDBJ)

The DNA Data Bank of Japan (DDBJ) is a biological database that collects DNA sequences. It is located at the National Institute of Genetics (NIG) in the Shizuoka prefecture of Japan. It is also a member of the International Nucleotide Sequence Database Collaboration or INSDC. It exchanges its data with European Molecular Biology Laboratory at the European Bioinformatics Institute and with GenBank at the National Center for Biotechnology Information on a daily basis. Thus these three databanks contain the same data at any given time.
DDBJ began data bank activities in 1986 at NIG and remains the only nucleotide sequence data bank in Asia. Although DDBJ mainly receives its data from Japanese researchers, it can accept data from contributors from any other country. DDBJ is primarily funded by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). DDBJ has an international advisory committee which consists of nine members, 3 members each from Europe, US, and Japan. This committee advises DDBJ about its maintenance, management and future plans once a year. Apart from this DDBJ also has an international collaborative committee which advises on various technical issues related to international collaboration and consists of working-level participants.

DDBJ has been functioning as one of the international nucleotide sequence databases, including EMBL-Bank/EBI in Europe and GenBank/NCBI in the USA as the two other members. DDBJ/EMBL-Bank/GenBank collects the nucleotide sequence data experimentally determined, and constructs the database in accordance with the rule agreed with the three databanks.

The database is a collection of "entry" which is the unit of the data. Each entry includes nucleotide sequence and the information of submitters, references, source organisms, and the biological nature such as gene function and other property of the sequence, etc.
The database also includes the data from Japan Patent Office (JPO), European Patent Office (EPO), United States Patent and Trademark Office (USPTO), and Korean Intellectual Property Office (KIPO). Nucleotide sequence records organismic evolution more directly than other biological meterials and thus is invaluable not only for research in life sciences but also human welfare in general. The database is, so to speak, a common treasure of human beings. The database is accessible freely online to anyone in the world.

Courtesy: DNA Databank of Japan, Japan

Dept. of MBBT Half Yearly Publications

Publications in 2011 (January to July)

1. Genome size determination and RAPDd analysis of four edible aroids of North East India. Jyoti Prasad Saikia and Bolin Kumar Konwar (2011). IIOAB Journal.

2. Biochemical composition and bioactivity of four edible aroids. Jyoti Prasad Saikia and Bolin Kumar Konwar (2011). Journal of Root Crops.

3. Production and physiochemical characterization of a biosurfactant produced by Pseudomonas aeruginosa OBP1 isolated from the petroleum sludge. P. Bharali and B. K. Konwar (2011). Applied Biochem. & Biotech.

4. Crude biosurfactant from thermophilic Alcaligenes faecalis: Feasibility in petro-spill biorem ediation. P. Bharali, S. Das, B. K. Konwar and A. J. Thakur (2011). Int J Biodeterioration & Biodegradation.

5. Enhancing the stability of colloidal silver nanoparticles using polyhydroxyalkanoates (PHA) from Bacillus circulans (MTCC 8167) isolated from crude oil contaminated soil Colloids and Surfaces B: Biointerfaces. Pinkee Phukan, J. P.Saikia and B. K. Konwar (2011). Colloids and Surfaces B: Biointerfaces.

6. Synthesis of silver polystyrene nanocomposite particles using water in supercritical carbon dioxide medium and its antimicrobial activity. I.R. Kamrupi, P. Phukon, B.K. Konwer and S.K. Dolui (2011). The Journal of Supercritical Fluids.

7. Enhanced visible light photocatalytic disinfection of gram negative, pathogenic Escherichia coli bacteria with Ag/ TiV oxide-nanoparticle. Nair, R.G., Roy, J.K., Samdarshi, S.K., Mukherjee, A.K.(2011). Colloids and Surface B: Biointerfaces.

8. An acidic phospholipase A2 (RVVA-PLA2-I) purified from Daboia russelli venom exerts its anticoagulant activity by enzymatic hydrolysis of plasma phospholipids and by non-enzymatic inhibition of factor Xa in a phospholipids/Ca2+ independent manner. Saikia D., Thakur, R., Mukherjee, A. K. (2011) Toxicon.

9. Optimization of production of an oxidant and detergent-stable alkaline β-keratinase from Brevibacillus sp. strain AS-S10-II: application of enzyme in laundry detergent formulations and in leather industry. Rai, S. K., Mukherjee, A. K. (2011). Biochemical Engineering Journal.

10. A Statistical approach for the enhanced production of alkaline protease showing fibrinolytic activity from a newly isolated Gram-negative Bacillus sp. strain AS- S20-I . Mukherjee, A. K., Rai, S.K. (2011). New Biotechnology.

Bio-Software's: AutoDock

AutoDock is a suite of automated docking tools. It is designed to predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure. The current distributions of AutoDock consist of two generations of software: AutoDock 4 and AutoDock Vina.

AutoDock 4

AutoDock 4 actually consists of two main programs: (i) autodock and (ii) autogrid. Autodock performs the docking of the ligand to a set of grids describing the target protein; Autogrid pre-calculates these grids. In addition to using them for docking, the atomic affinity grids can be visualised. This can help, for example, to guide organic synthetic chemists design better binders.

AutoDock 4.0 also allows sidechains in the macromolecule to be flexible. As before, rigid docking is blindingly fast, and high-quality flexible docking can be done in around a minute. Up to 40,000 rigid dockings can be done in a day on one cpu. AutoDock 4.0 now has a free-energy scoring function that is based on a linear regression analysis, the AMBER force field, and an even larger set of diverse protein-ligand complexes with known inhibiton constants.

AutoDock Vina

AutoDock Vina is a new generation of docking software from the Molecular Graphics Lab. It achieves significant improvements in the average accuracy of the binding mode predictions, while also being up to two orders of magnitude faster than AutoDock 4. AutoDock Vina does not require choosing atom types and pre-calculating grid maps for them. Instead, it calculates the grids internally, for the atom types that are needed, and it does this virtually instantly. Because the scoring functions used by AutoDock 4 and AutoDock Vina are different and inexact, on any given problem, either program may provide a better result.

AutoDock has applications in: (i) X-ray crystallography; (ii) structure-based drug design; (iii) lead optimization; (iv) virtual screening (HTS); (v) combinatorial library design; (vi) protein-protein docking; (vi) chemical mechanism studies.

AutoDock has been widely-used and there are many examples of its successful application in the literature. AutoDock is one of the most cited docking software . It is very fast, provides high quality predictions of ligand conformations, and good correlations between predicted inhibition constants and experimental ones. AutoDock has also been shown to be useful in blind docking, where the location of the binding site is not known. Plus,
AutoDock 4 is free and is available under the GNU General Public License. AutoDock Vina is available under the Apache license, allowing commercial and non-commercial use and redistribution.

Courtesy: The Scripps Research Institute.

Algorithms: Needleman-Wunsch Algorithm

The Needleman–Wunsch algorithm is a non linear global optimization method that was developed for amino acid sequence alignment in proteins. The algorithm performs a global alignment on two sequences (called A and B here). It is commonly used in bioinformatics to align protein or nucleotide sequences. The algorithm was published in 1970 by Saul B. Needleman and Christian D. Wunsch.

The Needleman–Wunsch algorithm is an example of dynamic programming, and was the first application of dynamic programming to biological sequence comparison.

Scores for aligned characters are specified by a similarity matrix. Here, $S(a,b)$ is the similarity of characters a and b. It uses a linear gap penalty, here called d.

For example, if the similarity matrix were

	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

Then the alignment

```
AGACTAGTTAC
CGA---GACGT
```

with a gap penalty of -5, would have the following score:

$$S(A,C)+S(G,G)+S(A,A)+(3\times d)+S(G,G)+S(T,A)+S(T,C)+S(A,G)+S(C,T)$$

Needleman and Wunsch described their algorithm explicitly for the case when the alignment is penalized solely by the matches and mismatches, and gaps have no penalty (d=0). In modern terminology, "Needleman-Wunsch" refers to a global alignment algorithm that takes quadratic time for a linear or affine gap penalty.

Bio-Quiz - 04

1. What was the first restriction enzyme to be characterized?
A) Sau 96
B) Eco RI
C) Hind III

2. Which of the following genes are responsible for nitrogen fixation?
(A) Nif genes
(B) Nod genes
(C) Nr genes

3. Which of these type of biological movements is mediated by Tubulin and Dynein?
(A) Beating of intestinal microvilli
(B) Cleavage of cells in mitosis
(C) Movement of chromosomes in mitosis

4. Which bacteria is the cause of whooping cough?
(A) Bordetella pertussis
(B) Brucella abortus
(C) Micrococcus pyogenes

5. What was AZT originally used for?
(A) Treatment of AIDS
(B) Treatment of cancer
(C) Treatment of warts

Answers: 1-B ; 2-A ; 3-C ; 4-A ; 5–B

For Suggestions and Contributions Please contact :

Prof. B.K. Konwar, Coordinator BIF

Dept of Molecular Biology & Biotechnology,

Tezpur University, Tezpur; email: bkkon@tezu.ernet.in

Mr. Salam Pradeep, Research Associate BIF

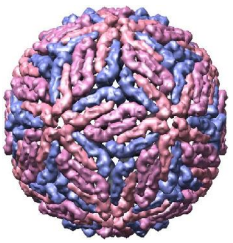
Dept of Molecular Biology & Biotechnology,

Tezpur University, Tezpur; email: salampradeep@gmail.com

Protein in Digest - Dengue Virus

Dengue virus is a major threat to health in tropical countries around the world. It is limited primarily to the tropics because it is transmitted by a tropical mosquito of the Aedes genus, principally *A. aegypti.* Most infected people experience dengue fever, with terrible headaches and fever and rashes that last a week or two. In some cases, however, the virus weakens the circulatory system and can lead to deadly hemorrhaging. Researchers are now actively studying the virus to try to develop drugs to cure infection, and vaccines to block infection before it starts.

Dengue virus is a small virus that carries a single strand of RNA as its genome. The genome encodes only ten proteins. Three of these are structural proteins that form the coat of the virus and deliver the RNA to target cells, and seven of them are nonstructural proteins that orchestrate the production of new viruses once the virus gets inside the cell. The outermost structural protein, termed the envelope protein, is shown here from PDB entry 1k4r. The virus is enveloped with a lipid membrane, and 180 identical copies of the envelope protein are attached to the surface of the membrane by a short transmembrane segment. The job of the envelope protein is to attach to a cell surface and begin the process of infection.



In the infectious form of the virus, the envelope protein lays flat on the surface of the virus, forming a smooth coat with icosahedral symmetry. However, when the virus is carried into the cell and into lysozomes, the acidic environment causes the protein to snap into a different shape, assembling into trimeric spike. Several hydrophobic amino acids at the tip of this spike, insert into the lysozomal membrane and cause the virus membrane to fuse with lysozome. This releases the RNA into the cell and infection starts. The hemagglutinin protein on the surface of influenza virus plays a similar role, but the two proteins use entirely different mechanisms to perform a similar task.

A dengue vaccine has proven difficult to develop, in part because there are four major subtypes of dengue virus, each with slightly different viral proteins. Many researchers currently believe that the deadly dengue hemorrhagic disease is caused when a person is infected with one subtype, and then infected later by a second subtype. The antibodies, and immunity, gained from the first infection appear to assist with the infection by the second subtype, instead of providing a general immunity to all subtypes. This means that an effective vaccine will have to stimulate protective antibodies against all four types at once, a feat that has not yet been achieved.

Courtesy: RCSB, Protein Data Bank.