



Bioinformation up to Date

(Bioinformatics Infrastructure Facility, Biotechnology Division)
North-East Institute of Science & Technology
 Jorhat - 785006, Assam

Contents	Cover Story																						
<table border="1"> <tr> <td style="width: 80%;">Cover Story</td> <td style="text-align: right;">1</td> </tr> <tr> <td>Special Interest</td> <td style="text-align: right;">2</td> </tr> <tr> <td>Proteomics</td> <td style="text-align: right;">2</td> </tr> <tr> <td>Genomics</td> <td style="text-align: right;">2</td> </tr> <tr> <td>Software Mania</td> <td style="text-align: right;">3</td> </tr> <tr> <td>Bio Server</td> <td style="text-align: right;">3</td> </tr> <tr> <td>Bioinfo Quiz</td> <td style="text-align: right;">3</td> </tr> <tr> <td>Computational Chemistry</td> <td style="text-align: right;">4</td> </tr> <tr> <td>Molecule of the Month</td> <td style="text-align: right;">4</td> </tr> <tr> <td>Bioinfo Animator</td> <td style="text-align: right;">4</td> </tr> <tr> <td>Contact Us</td> <td style="text-align: right;">4</td> </tr> </table>	Cover Story	1	Special Interest	2	Proteomics	2	Genomics	2	Software Mania	3	Bio Server	3	Bioinfo Quiz	3	Computational Chemistry	4	Molecule of the Month	4	Bioinfo Animator	4	Contact Us	4	<h3 style="text-align: center;">Report on Short Term Training Course on Basics of Bioinformatics</h3> <p>A three days short term training course on Basics of Bioinformatics was held on 24th March - 26th March, 2009 at the Bioinformatics Infrastructure Facility, Biotechnology Division, North-East Institute of Science and Technology, Jorhat, Assam. A total of 13 participants from various institute such as Central Silk Board, Jorhat, Tea Research Association, Jorhat, Defense Research Laboratory, Tezpur, Dibrugarh University, Gauhati University etc were present. Each of the participants was provided with a folder which contains bioinformatics software's CD, a course material on bioinformatics, other necessary accessories.</p> <p>The training course was inaugurated by Dr. P.G. Rao, Director North-East Institute of Science & Technology, Jorhat.</p> <p>On 24th March, 2009, the first session commenced with a Lecture on "Basics of Bioinformatics" by Mr. Salam Pradeep Singh, CSIR Diamond Jubilee Research Intern, Biotechnology Division, North-East Institute of Science & Technology, Jorhat.</p> <p>The second session was on demonstration of various biological sequence databases & information retrieval and demonstration on various sequence analysis & alignment software's and servers by Mr. Salam Pradeep Singh. This was followed by a discussion session with the participants.</p> <p>The afternoon program me was on hands on practical session, each participant was allotted a single system. Hands on training were given on accessing Biological Sequence Databases such as NCBI Genbank, EMBL Nucleotide DB, DDBJ, SwissProt, UniProt etc and Sequence Analysis & Alignment software's such as CLC sequence viewer, EBI ClustalW server, NCBI BLAST for sequence similarity search etc.</p> <p>On the second day i.e. 25th March, 2009, there was lecture on Protein Structures by Associate Professor Dr. M.K. Modi, Dept. of Biotechnology, Assam Agricultural University, Jorhat. This was followed by demonstration on Protein Structure Databases and protein three dimensional prediction servers by Shri Salam Pradeep Singh.</p> <p>The afternoon session was on hands on practical session. Hands on training was given Protein Structure Data downloads and Protein 3D structure Visualization with different molecular visualization software's. The participants were also trained how to analyze the protein structure using various software's.</p> <p>On the third day, i.e. 26th March, 2009, there was lecture on Molecular Phylogenetics by Dr. M.A. Laskar, Coordinator BIF, St. Anthony's College, Shillong.</p> <p>After his lecture Dr. M.A Laskar demonstrated various software on molecular Phylogenetic Analysis After the software demonstration by Dr. M.A. Laskar, there was software demonstration on molecular biology analysis such as PCR primer designing, restriction etc analysis, virtual electrophoresis plot by Mr. Salam Pradeep.</p> <p>The afternoon session was on hands on training on molecular biology software's for PCR primer designing, restriction site analysis and virtual electrophoresis. This was followed by the valedictory function.</p>
Cover Story	1																						
Special Interest	2																						
Proteomics	2																						
Genomics	2																						
Software Mania	3																						
Bio Server	3																						
Bioinfo Quiz	3																						
Computational Chemistry	4																						
Molecule of the Month	4																						
Bioinfo Animator	4																						
Contact Us	4																						
<p>Adviser: Dr. P.G. Rao</p> <p>Editors: Salam Pradeep Singh Dr. R.L. Bezbaruah</p>																							
<p>Important Events</p> <ol style="list-style-type: none"> 1. 5th International Conference on "Biopesticides". 26-30 April 2009 @ Indian Habitat Center (IHC) Lodhi Road, New Delhi, India. 2. A web enabled application for submission of project proposals to the Department of Biotechnology (DBT) is available at http://dbtpms.nic.in . All PI's associated with DBT are being informed about their User ID / Password to access the system. 																							

Readers Contribution

Personal Experience on Short Term Training Course on Basics of Bioinformatics

I got the chance to attend a three days short term training on basics of bioinformatics that was held in Bioinformatics Infrastructure facility, Biotechnology Division, North-East Institute of Science & Technology, Jorhat. The training was very helpful and we got to know about many bioinformatics software that can be easily availed from the internet. These different software's have numerous applications like sequence alignment and similarity search, phylogenetic analysis, primer designing, protein structure prediction, molecular modeling etc. and they can be applied to assist our research work. Also we got to know about many paid software that have additional advantages.

The lectures delivered by Prof. M.K. Modi and Prof. M.A. Laskar were very useful. Particularly Prof. M.A. Laskar gave us a very detailed account of how to use NTSYS PC, a paid software for molecular phenetic studies. A lot of information was disseminated on how to construct a phylogenetic tree from wet lab results and by using NTSYS PC and Phylip software. A very important term used in molecular phonetic studies called bootstrap analysis was explained to us. Bootstrap analysis and statistical analysis can be said to be synonymous.

The hands on session demonstrated by Salam Pradeep Singh were very helpful and informative and we got a chance to practice and use all the bioinformatics software.

Contributed by: Miss Susmita Singh, Senior Research Fellow, Biotechnology Division

Proteomics

SUPERFAMILY

SUPERFAMILY is a database of structural and functional protein annotations for all completely sequenced organisms.

A domain is the smallest unit of evolution; a large protein can be split into smaller domains. Domains can occur by themselves or in combination with other domains. A superfamily groups together domains of different families which have a common evolutionary ancestor based on structural, functional and evolutionary data. The SUPERFAMILY web site and database provides protein domain assignments, at the SCOP 'superfamily' and 'family' levels, for the predicted protein sequences in over 900 organisms.

SUPERFAMILY domain assignments are generated using an expert curated set of profile hidden Markov models. All models and structural assignments are available for browsing and download.

Major Important Features:

Sequence search: Submit your protein, or DNA, sequence for superfamily and family level classification.

Keyword search: Search for superfamily names, family names, species names, sequence IDs, SCOP IDs, PDB IDs and hidden Markov model IDs.

Domain assignments: Domain assignments, alignments and architectures for completely sequenced eukaryotic, and prokaryotic organisms.

Comparative genomics tools: Browse unusual superfamilies & families, adjacent domain pair lists & graphs, unique domain pairs, combinations, & domain distribution across taxonomic kingdoms for each organism.

Genome statistics: For each genome: number of sequences, number & percentage of sequences with assignment, percentage total sequence coverage, number of domains & superfamilies assigned, number of families assigned, average superfamily size, percentage produced by duplication, average sequence length & matched etc.

Superfamily annotation: InterPro have added abstracts for 1,052 superfamilies, and 763 superfamilies have some Gene Ontology (GO) annotation.

Functional annotation: Functional annotation of SCOP superfamilies.

Phylogenetic trees: Genome combinations, or specific clades, can be displayed as individual trees. The trees are based on protein domain architecture data for all genomes in SUPERFAMILY, and are generated using heuristic parsimony methods.

Similar domain architectures: Finds the 10 domain architectures which are most similar to a domain architecture of interest.

Courtesy: Medical Research Council, U.K.

Genomics

National Microbial Pathogen Data Resource - NMPDR

The NMPDR provides curated annotations in an environment for comparative analysis of genomes and biological subsystems, with an emphasis on the food-borne pathogens *Campylobacter*, *Listeria*, *Staphylococcus*, *Streptococcus*, and *Vibrio*; as well as the STD pathogens *Chlamydiae*, *Haemophilus*, *Mycoplasma*, *Neisseria*, *Treponema*, and *Ureaplasma*.

This edition of the NMPDR includes 47 archaeal, 725 bacterial, and 29 eukaryal genomes with 3,257,053 genetic features, of which 1,302,831 are in FIGfams curated using 616 active subsystems.

Some Important Features:

1. Annotate

a) RAST: Rapid Annotation Server to electronically annotate large Genomes.

b) MG-RAST: Metagenomics Rapid Annotation Server, a variant of the RAST for metagenomes.

2. Compare

a) Annotations: Display all the annotations submitted to the Annotation Clearinghouse for a selected genome.

b) Proteomes: Display a table of Bidirectional Best Hits between selected genomes.

c) Subsystems: Compare the metabolic pathways used by 2 organisms.

d) Signature Genes: Find genes that are common among selected genomes or differentiate two sets of genomes.

e) FIGfams: Find the protein family of a FASTA Format amino acid sequence.

f) KEGG maps: View the KEGG diagrams for a particular genome.

3. Search

a) BLAST or Scan: Search for protein or DNA sequences using BLAST or our advanced pattern-scanning tool.

b) Genes: Search for genes in selected genomes using keywords, a specific subsystem, or both.

c) Organisms: View a summary of NMPDR core organisms and a list of supporting organisms.

d) Subsystems: View the tree of subsystems, and do keyword searches in individual subsystems or subsystem categories.

e) Protein Targets: Search for genes using a combination of useful criteria.

Courtesy: NMPDR, Chicago

Software Mania

FlexX



FlexX is a computer program for predicting protein-ligand interactions. For a given protein and a ligand, FlexX predicts the geometry of the complex as well as an estimate for the strength of binding. In this first version of FlexX, the protein is assumed to be rigid. Thus, the protein must be given in a conformation which is similar to the bound state. The docking algorithm in FlexX works without manual intervention. FlexX is ideal for interactive work on protein-ligand complexes as well as for screening a larger set of ligands in order to find new leads for drug design. In summary, FlexX can be useful in the following situation:

We have a good three-dimensional model of the protein and you know the location of the active site. We have a set of ligands and we want to know whether and how each of them binds to your protein model.

FlexX originates from research as part of the RELIWE1 and RELIMO2 projects at the German National Research Center for Information Technology (GMD), Institute for Algorithms and Scientific Computing (SCAI).

FlexX's workflow consists of 4 major steps:

- (1) Protein Preparation (2) Ligand and Library Composition (3) Docking (4) Analysis

Courtesy: BioSolveIT GmbH

Bio Servers

PlantProm DB - Database of Plant Promoter Sequence

PlantProm DB is an annotated, non-redundant collection of proximal promoter sequences for RNA polymerase II with experimentally determined transcription start site(s), TSS, from various plant species. It was developed by Softberry in collaboration with Department of Computer Science at Royal Holloway, University of London. Current release of PlantProm DB contains 305 entries, including 71, 220 and 14 promoters from monocot, dicot and other plants, respectively.

There is experimental evidence of the TSS position(s) of the gene, published in the literature. For genes with multiple TSSs, the nearest to the CDS start position is taken, if no additional information on the predominance of one of them is available (positions of other TSSs are given in the name line of the sequence written in the FASTA format).

The length of known promoter sequence upstream of chosen TSS is 200 bp or more; all stored promoter sequences are the same length, 251 bp, where the position 201 corresponds to the TSS, i.e. collected sequences occupy the region [-200 : +51], with the TSS in the position +1, and, thus, present proximal promoters mentioned above.

Each entry corresponds to a gene mapped on genomic sequence. Various alleles of a single gene are presented in the database by a single entry. Genes with more than one non-allelic copy in the genome, as well as paralogous genes, are treated as different entries.

PlantProm DB provides the following information.

1. DNA sequence of 305 promoter regions [-200:+51], with TSS on the fixed position +201, from various plant species, in the FASTA format, including:

- a) 71 promoters of monocots, b) 220 promoters of dicots, c) 14 promoters from other plants,
d) 175 TATA promoters, consisting of 41 monocot, 131 dicot & 3 other plant species sequences, respectively. e) 130 TATA-less promoters, consisting of 30 monocot, 89 dicot and 11 other plant species sequences, respectively.

2. Taxonomic and promoter type classification of promoters, including:

- a) List of species represented in the PlantProm DB,
b) List of genes/gene products and promoter types represented in the PlantProm DB.

3. Nucleotide Frequency Matrices for canonical promoter elements (TATA-box, CCAAT- box, and TSS-motif or Initiator element, Inr), including:

- a) TATA-matrices for various promoter collections,
b) CCAAT-matrices for various promoter collections, c) TSS-motif-matrices for various promoter collections.

4. Location of TATA-boxes in some promoters collections mentioned above, including:

- a) 171 unrelated promoters from various plant species, b) 128 unrelated promoters from dicot plants.

5. Location of CCAAT-boxes in some promoters collections mentioned above, including:

- a) 131 unrelated promoters of both (TATA and TATA-less) types from various plant species,
b) 71 unrelated TATA promoters from various plant species,
c) 60 unrelated TATA-less promoters from various plant species.

6. Location of TSS-motifs in some promoters collections mentioned above, including:

- a) 70 unrelated promoters of both (TATA and TATA-less) types from monocot plants,
b) 217 unrelated promoters of both (TATA and TATA-less) types from dicot plants,
c) 171 unrelated TATA promoters from various plant species, d) 130 unrelated TATA-less promoters from various plant species.

Courtesy: Softberry, Inc, New York

Bioinfo Quiz - 011

1. Occasionally the codon UGA can code for the amino acid:

- a) Cysteine
b) Methionine
c) Selenocysteine

2. Which of these enzymes contains a Zinc (Zn) ion?

- a) Carboxypeptidase A
b) Phosphorylase B kinase
c) Tyrosine hydroxylase

3. The main damaging effect of UV radiation on DNA is:

- a) Depurination
b) Thymine dimers formation
c) Single strand break

4. What reagent is used to measure the number of thiol groups in a protein?

- a) Ellman's reagent
b) Mercaptoethanol
c) Ninhydrin

5. The approximate length of the H bonds in helical DNA A-T or G-C base pairs is?

- a) 1.5 Angstroms
b) 2.0 Angstroms
c) 3.0 Angstroms

6. Angiotensin converting enzyme requires for activity:

- a) NADH
b) Zinc ions
c) Magnesium ions and glutamine

Answers on Page 4

Computational Chemistry:

AMBER

AMBER (an acronym for Assisted Model Building and Energy Refinement) is a family of force fields for molecular dynamics of biomolecules originally developed by the late Peter Kollman's group at the University of California, San Francisco. AMBER is also the name for the molecular dynamics simulation package that implements these force fields. It is maintained by an active collaboration between David Case at Rutgers University, Tom Cheatham at the University of Utah, Tom Darden at NIEHS, Ken Merz at Florida, Carlos Simmerling at Stony Brook University, Ray Luo at UC Irvine, and Junmei Wang at Encysive Pharmaceuticals.

The term "AMBER force field" generally refers to the functional form used by the family of AMBER force fields. This form includes a number of parameters; each member of the family of AMBER force fields provides values for these parameters and has its own name.

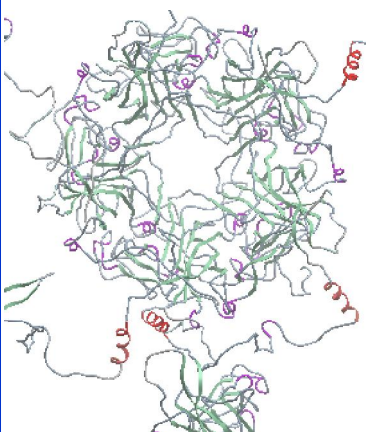
The functional form of the AMBER force field is shown below:

$$V(r^N) = \sum_{\text{bonds}} \frac{1}{2} k_b (l - l_0)^2 + \sum_{\text{angles}} k_a (\theta - \theta_0)^2 + \sum_{\text{torsions}} \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^N \left\{ \epsilon_{i,j} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

Molecule of the Month

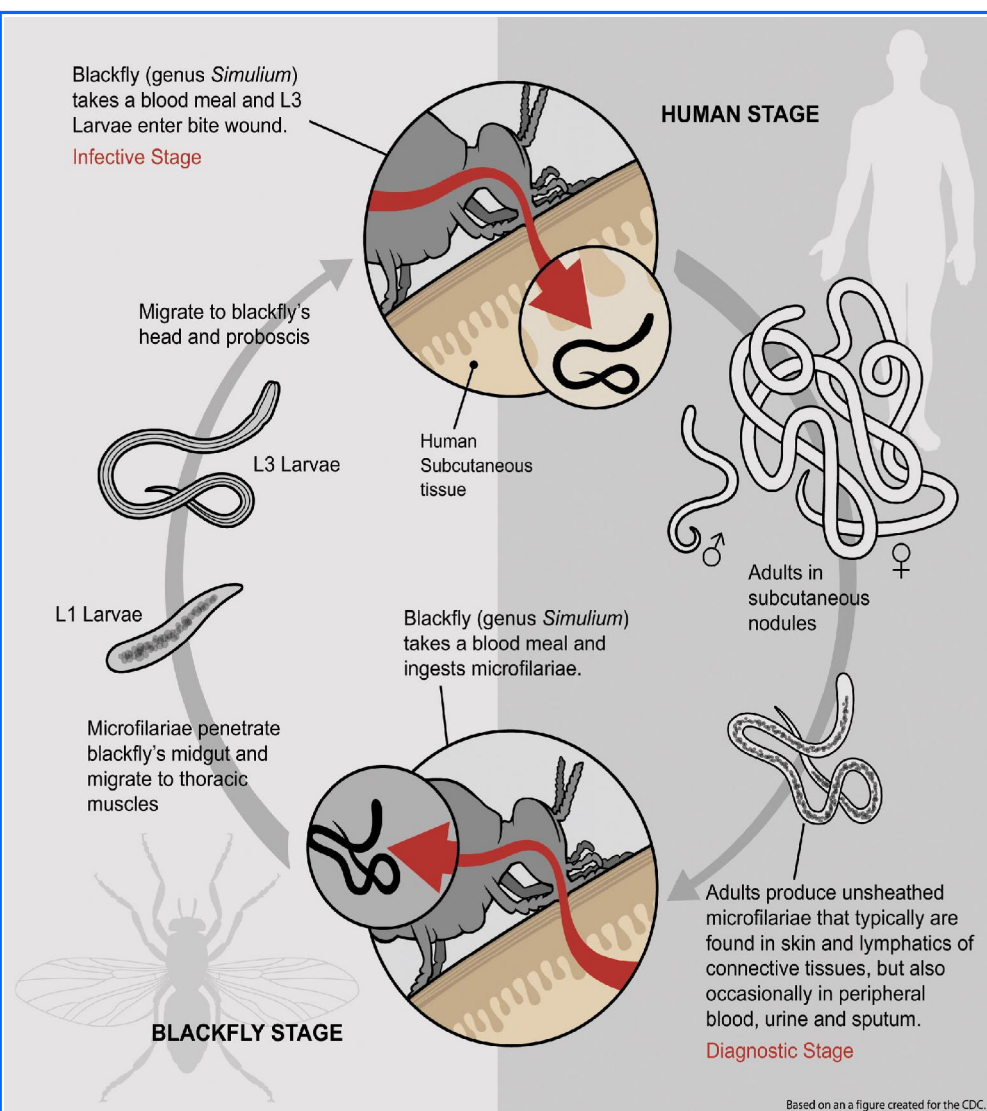
Simian Virus 40

Simian virus 40 is an example of how simple a virus can be and still perform its deadly job. Viruses are tiny machines with a single purpose to reproduce themselves. They enter cells and hijack their synthetic machinery, forcing them to create new viruses. SV40 does this with very little molecular machinery. It is enclosed by a spherical capsid composed of 360 copies of one protein, and a few copies of two others. This capsid is just big enough to enclose a small circle of DNA 5243 nucleotides long, which contains the barest minimum of information needed to get into the cell and make new viruses.



Molecular Data

PDB ID	: 1SVA
Amino acids	: 2057
Atoms	: 15989
Exp. Method	: X-ray
Chains	: 6
Structure Wt.	: 238915.81



Bioinfy Animator:- The life cycle of *Onchocerca volvulus*, a parasitic worm which causes river blindness (world's third leading infectious cause of blindness)

Courtesy: Centers for Disease Control and Prevention, Georgia, USA.

Please contribute, contact:

Salam Pradeep; CSIR D.J. Res. Intn;
Email: salampradeep@gmail.com.

Bioinfy Quiz

011
Answers

1 - c ; 2 - a ; 3 - b ;
4 - a ; 5 - c ; 6 - b