



Bioinformation up to Date

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

Contents

Cover Story	1
Special Interests	1
Computational Chemistry	2
Proteomics	2
Genomics	2
Software Mania	3
Bio Server	3
Bioinfy Quiz	3
Computers for Biologist	4
Molecule of the Month	4
Bioinfy Animator	4
Contact Us	4

Adviser:

Dr. P.G. Rao

Editors:

Salam Pradeep Singh

Dr. R.L. Bezbaruah

BIF Upcoming Events

1. Training on “**Bioinformatics – General concepts and applications**” @ BIF, Department of Biotechnology, Mizoram University from March 26th – 27th, 2009.

2. Training course on “**Application of Bioinformatics Tools in Biotechnology**” @ BIF, Gauhati University, Gawahati from 25th - 27th March, 2009.

Cover Story

Short Term training Course on Basics of Bioinformatics

Bioinformatics Infrastructure Facility, Biotechnology Division, North-East Institute of Science & Technology, Jorhat is organizing a Short Term Training Course on Basics of Bioinformatics to be held from 24th –26th March, 2009 for the first time in the newly inaugurated BIF room. The training program is supported by the Department of Biotechnology, Government of India, New Delhi.

The sole purpose of the training program is to make the life science students, research scholars and scientists of North-Eastern region aware of Bioinformatics software’s and computational biology software’s and biological databases which is available for public and how it can be applied in their day to day research activities.

The training course will include hands on practical session on how to retrieve data from biological databases and how to analyze the data. A number of bioinformatics software and computational biology software will be demonstrated and will also make the participants allow to practice the software’s. Apart from this software’s CD’s and course material will also be provided to the participants.

Special Interests



ImmunoGrid

The European Virtual Human Immune System Project

ImmunoGrid

ImmunoGrid is a STREP project funded by European Commission which started on February 1, 2006. The primary aim of them is the development and implementation of a Grid-based simulator of human immune system for support of clinically relevant applications, such as vaccine development and optimization of immunotherapies.

Immune system is a complex adaptive learning system evolved to maintain the healthy state of the organism and to protect against infectious diseases and cancers. It operates at multiple levels: molecule, cell, tissue, organ, and population. It has immense diversity due to its combinatorial nature.

Computational models are ideal for bridging these gaps. Modeling the immune system is a formidable challenge because of large number of components to be considered. There are some 1011 T cells and a similar number of B cells in human body. There are more than 1015 possible combinations of major histocompatibility complex proteins (human leukocyte antigens in human) which present targets of immune responses to the immune system. The revolution in information technology has ensured that available computational resources can deal with the systems of large complexity.

The emergence of the Grid Computing enables modeling human immune system at a natural scale.

Main aims of the ImmunoGrid are:

- Standardization of immunological concepts and related bioinformatics tools and resources
- Combining data, tools and resources to develop simulator and create models of human immune system
- Develop pre-clinical applications of the simulator testing
- Disseminate results and tools to researchers and clinicians

Computational Chemistry:

GAMESS

GAMESS is a computational chemistry software program that stands for General Atomic and Molecular Electronic Structure System. The original code split in 1981 into GAMESS (US) and GAMESS (UK) variants, which now differ significantly. GAMESS is maintained by the members of the Gordon research group at Iowa State University, USA.

GAMESS can perform a number of general computational chemistry calculations, including Hartree-Fock, Density functional theory (DFT), Generalized Valence Bond (GVB), and Multi-configurational self-consistent field (MCSCF). Correlation corrections after these SCF calculations can be estimated by configuration interaction (CI), second order Møller-Plesset perturbation theory, and coupled cluster theory. Solvent effect can be considered using discrete effective fragment potentials or continuum models (such as PCM). Relativistic corrections can be calculated, including third order Douglas-Kroll scalar terms.

While the program does not perform molecular mechanics, it can be interfaced with the TINKER code for molecular mechanics to do mixed molecular mechanics/quantum mechanics calculations. The Fragment Molecular Orbital method can be used to treat large systems, by dividing them into fragments.

It can also be interfaced with the valence bond VB2000 and XMVB programs and the Natural Bond Orbital (NBO) population analysis program.

Courtesy: GAMESS, USA

Proteomics

EXPASY - ProtParam

The ProtParam is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity.

ProtParam computes various physio-chemical properties that can be deduced from a protein sequence. No additional information is required about the protein under consideration. The protein can either be specified as a Swiss-Prot/TrEMBL accession number or ID, or in the form of a raw sequence. White spaces and numbers are ignored. If we provide the accession number of a Swiss-Prot/ TrEMBL entry, we will be prompted with an intermediary page that allows us to select the portion of the sequence on which we would like to perform the analysis.

The choice includes a selection of mature chains of peptides and domains from the Swiss-Prot feature table which can be chosen by clicking on the positions, as well as the possibility to enter start and end position in two boxes. By default (i.e. if we leave the two boxes empty) the complete sequence will be analyzed.

All computations performed by ProtParam are based on either compositional data, or on the N-terminal amino acid.

The parameters computed by ProtParam include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). Molecular weight and theoretical pI are calculated as in Compute pI/Mw.

Courtesy: Swiss Institute of Bioinformatics, Geneva

Genomics

Cancer Immunome Database - SEREX

The SEREX project, sponsored by the Ludwig Institute for cancer Research, aims to document the repertoire of antigens eliciting an antibody response in cancer patients. The sequences deposited in this database were obtained by screening cDNA expression libraries using serum from cancer patients as probes, and sequencing individual reactive clones. The technique was developed by the group of Michael Pfreundschuh at Saarland University. SEREX entries can be retrieved based on multiple criteria, such as the SEREX sequence ID, the patient, the lab or contributor etc.

Search SEREX

Gene:

Clone:

Sequence ID:

Sequence pattern:

as MySQL regex

cDNA library from

Patient:

Tissue:

Classification:

Serum from

Patient:

Classification:

Contributed by

Contributor:

Laboratory:

List that

genes satisfy all the above criteria

clones satisfy any one of the criteria

sequences map to Human Genome

Courtesy: Genetics Society of America, USA

Software Mania

Lasergene



Lasergene is a comprehensive software for DNA and protein sequence analysis, contig assembly and sequence project management. DNASTAR Lasergene software consists of an integrated suite of seven modules that can be purchased in any combination. The modules of Lasergene are:

- » SeqBuilder - Visualization and sequence
- » SeqBuilder - Primer Design
- » SeqMan Pro - Sequence assembly and SNP discovery
- » SeqMan Pro - SNP Discovery and Reporting
- » MegAlign - Sequence alignment
- » PrimerSelect - Oligo primer
- » Protean - Protein structure analysis & prediction
- » GeneQuest - Gene finding
- » EditSeq - Utility for importing unusual file types

Lasergene has also the features that simplify Primer Design, SNP analysis and protein structural analysis. New primer designing capabilities that have been added to SeqBuilder simplify mutagenesis and virtual cloning.

Bio Servers

CPH Model 2.0 Server

CPH Model 2.0 Server is a novel method that was developed for fold recognition/homology modeling, in which a large sequence database is iteratively searched to construct a sequence profile until a template can be found in a database of proteins with known structure. The method differs from the PDB-BLAST method in that a sequence profile is only made if a template is not readily found in the database of known structures. A sequence profile is subsequently made for the template, using the same number of PSI-BLAST iterations that were used to identify it. Query and template sequences are subsequently aligned using a score based on profile-profile comparisons. The alignment score is modified so as to ensure unreliable parts of the alignment is discarded.

The output is divided into the following sections:

Query sequence: In this section the query sequence that you submitted are shown in fasta format.

Searching for template: The template for building the model is sought by iteratively building up a profile by aligning the query sequence to a nonredundant database of protein sequences and then searching a database of proteins with known structure to find a suitable template for making a model.

Retrieving template: In this section the pdb entry name and the chain identifier are listed for the template that are used to construct the model.

Making profile-profile alignment: In this section the score from the profile profile alignment (in bits) and the percentage sequence identity between query and template are shown together with the alignment in "Blast-like" format.

Modeling: By clicking on the link "model.pdb" you can download the coordinates in pdb format to your own computer.

PDB3D: If we have an java enabled browser the C-alpha trace of the model will be shown.

SUBMISSION

Paste a single sequence or several sequences in **FASTA** format into the field below:

PASTE A PROTEIN SEQUENCE

Submit a file in **FASTA** format directly from your local disk:

Restrictions:

Only one sequence per submission with not less than 15 and not more than 4,000 amino acids.

Confidentiality:

The sequences are kept confidential and will be deleted after processing.

Bioinfo Quiz - 010

1. Coomassie Blue stains protein by reacting with:

- A) Arginine residues
- B) Free C-terminal
- C) Peptide bonds

2. Which of the following histones is the most conserved between species?

- A) Histone H1
- B) Histone H3
- C) Histone H4

3. Where are the ran proteins localized in the cell?

- A) Endoplasmic reticulum
- B) Golgi sacki
- C) Nucleus

4. How many different codons translate to leucine?

- A) 2
- B) 4
- C) 6

5. Who isolated the first Drosophila mutant in 1910?

- A) H.J Muller
- B) G. Mendel
- C) T.H. Morgan

Answers on Page 4

Smith-Waterman algorithm

The Smith-Waterman algorithm is a well-known algorithm for performing local sequence alignment; that is, for determining similar regions between two nucleotide or protein sequences. Instead of looking at the total sequence, the Smith-Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.

The algorithm was first proposed by Temple Smith and Michael Waterman in 1981. Like the Needleman-Wunsch algorithm, of which it is a variation, Smith-Waterman is a dynamic programming algorithm. As such, it has the desirable property that it is guaranteed to find the optimal local alignment with respect to the scoring system being used (which includes the substitution matrix and the gap-scoring scheme). The main difference to the Needleman-Wunsch algorithm is that negative scoring matrix cells are set to zero, which renders the (thus positively scoring) local alignments visible. Backtracing starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment.

A matrix H is built as follows:

$$H(i, 0) = 0, 0 \leq i \leq m$$

$$H(0, j) = 0, 0 \leq j \leq n$$

$$H(i, j) = \max \left\{ \begin{array}{l} H(i-1, j-1) + w(a_i, b_j) \quad \text{Match/Mismatch} \\ H(i-1, j) + w(a_i, -) \quad \text{Deletion} \\ H(i, j-1) + w(-, b_j) \quad \text{Insertion} \end{array} \right\}, 1 \leq i \leq m, 1 \leq j \leq n$$

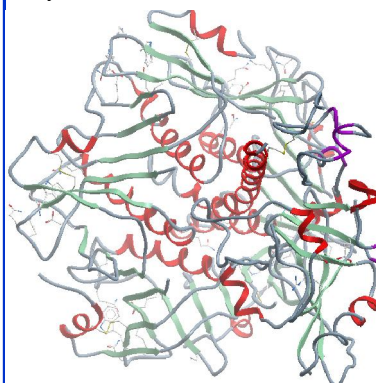
Where:

- a, b = Strings over the Alphabet Σ
- m = length(a)
- n = length(b)
- $H(i, j)$ - is the maximum **Similarity-Score** between the substring of a of length i , and the substring of b of length j
- $w(c, d)$, $c, d \in \Sigma \cup \{-\}$, w is the **gap-scoring scheme**

Molecule of the Month

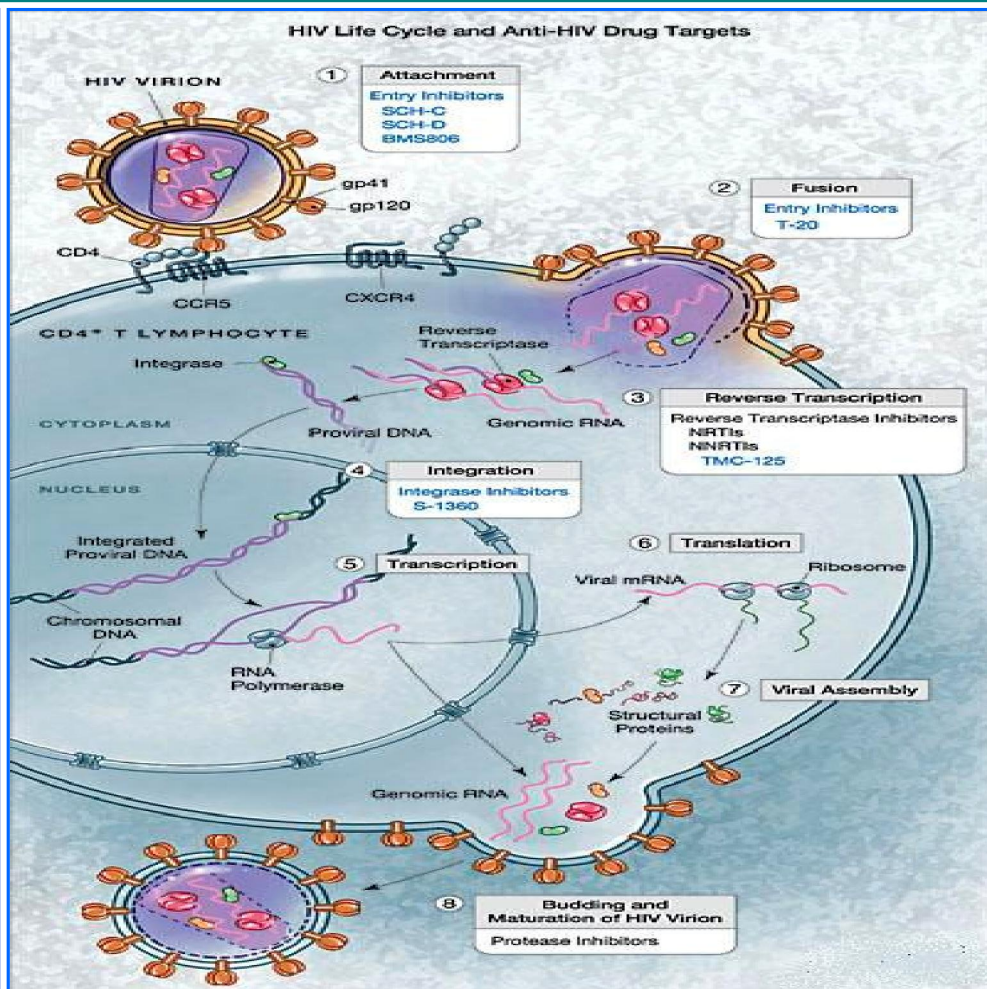
Cholera Toxin

Cholera toxin, structure shown below has a ring of five identical protein chains, which binds to carbohydrates on the surface of cells. This delivers the toxic part of the molecule, to the cell, where it can wreak its havoc. The catalytic portion of cholera toxin performs a single function: it seeks out the G proteins used for cellular signaling and attaches an ADP molecule to them. This converts the G-Protein into a permanently active state, so it sends a never-ending signal. This confuses the cell, and among other things, it begins to transport lots of water and sodium outwards. This floods the intestine, leading to life threatening dehydration.



Molecular Data

PDB ID	: 1XTC
Exp. Method	: X-ray
Chains	: Seven (7)
Classification	: Toxin



Bioinfy Animator:- Current anti-HIV drugs inhibit reverse transcriptase (nucleoside reverse transcriptase inhibitors [NRTIs] and nonnucleoside reverse transcriptase inhibitors [NNRTIs]) or protease. Anti-HIV drugs under development (blue) include agents that interfere with other steps in the HIV life cycle (entry inhibitors and integrase inhibitors) and a second-generation NNRTI.

Please contribute, contact:

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

Bioinfy Quiz 010

Answers

1 - A ; 2 - C ; 3 - C ;
4 - C ; 5 - C