



## Bioinformation up to Date

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

### Contents

Cover Story	1
Computational Chemistry	1
Genomics	2
Software Mania	3
Bio Server	3
Bioinfy Quiz	3
Proteomics	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

### Upcoming Events

**1. Short Term Training Course on “Biotechnology & Bioinformatics tools for Genomics & Proteomics”** from 23<sup>rd</sup> Sept - 26<sup>th</sup> Sept, 2009 organized by Bioinformatics Infrastructure Facility, Dibrugarh University, Dibrugarh. Includes Lectures, Demos & Hands on Session.

### Cover Story

#### Biotech Hubs & State Nodal Centres in North-East India

The Department of Biotechnology, Ministry of Science & Technology is inviting proposals from institutes/colleges/Universities of higher education of the North Eastern Region of India offering undergraduate and/or postgraduate courses in any branch of Biology/Life Science/ Biotechnology / Environmental Sciences/Chemistry /Physics/ Computer Science /Biomedical Sciences/ Agriculture/ Veterinary Sciences for establishment of Biotech Hubs under the special programme for the northeastern states.

R&D institutions in this region those are actively engaged in biological research will also be considered for this program. The broad purpose of the programme is to promote education and research in Biology/Life Science/Biotechnology and to attract brilliant young students to build their career in different fields of biological sciences/biotechnology. The support under this programme will be provided initially for a period of 3 years.

Under this programme DBT proposes to provide the following facilities to each Biotech Hub

- A biotechnology lab with basic set of equipments
- Support for site preparation/renovation
- A bioinformatics centre with min. of 5 computers with Internet connectivity
- Biotechnology electronic journal access facility
- Recurring budget for procurement of chemicals, glassware and for maintenance of Internet connectivity
- Support for Organizing Trainings/Workshops for teachers
- Travel Support for Conference allowance, short term training in any Institute in India.
- Student Fellowships etc.

The Biotech Hubs shall offer the facilities for promotion of effective teaching in life sciences as biotechnology in the host and the neighboring institutions. Biotech Hubs will organize summer and winter courses for school students and biology teachers of secondary and senior secondary schools.

Each state will have a nodal centre in a leading institution. The nodal centre will train the coordinators of the institutional hubs and periodically organize trainings and workshops for students and faculties. The nodal centre will have linkage with all biotech hubs within the state and will assist & monitor their activities.

The Biotechnology Division of NEIST, Jorhat has already applied to step up to such a facility which would enrich the Biotechnology research.

## Computational Chemistry

### Gradient Descent

Gradient descent is a first-order optimization algorithm. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or the approximate gradient) of the function at the current point. If instead one takes steps proportional to the gradient, one approaches a local maximum of that function; the procedure is then known as gradient ascent.

Gradient descent is also known as steepest descent, or the method of steepest descent. When known as the latter, gradient descent should not be confused with the method of steepest descent for approximating integrals.

Gradient descent is based on the observation that if the real-valued function  $F(X)$  is defined and differentiable in a neighborhood of a point  $a$ , then  $F(X)$  decreases fastest if one goes from  $a$  in the direction of the negative gradient of

$$F \text{ at } a, - \nabla F(a). \text{ It follows that, if } b = a - \nabla F(a)$$

For  $V > 0$  a small enough number, then  $F(a) > F(b)$ . With this observation in mind, one starts with a guess  $X_0$  for a local minimum of  $F$ , and considers the sequence  $X_0, X_1, X_2 \dots$  such that

$$X_{n+1} = X_n - V \nabla F(X_n), n \geq 0.$$

We have,

$$F(X_0) > F(X_1) > F(X_2) \dots,$$

So hopefully the sequence  $(X_n)$  coverage to the desired local minimum. Note that the value of the step size  $V$  is allowed to change at every iteration.

This process is illustrated in the picture to the right. Here  $F$  is assumed to be defined on the plane, and that its graph has a bowl shape. The blue curves are the contour lines, that is, the regions on which the value of  $F$  is constant. A red arrow originating at a point shows the direction of the negative gradient at that point. Note that the (negative) gradient at a point is orthogonal to the contour line going through that point. We see that gradient descent leads us to the bottom of the bowl, that is, to the point where the value of the function  $F$  is minimal.

## Genomics

### NCBI Taxonomy

Organismal taxonomy is a powerful organizing principle in the study of biological systems. Inheritance, homology by common descent, and the conservation of sequence and structure in the determination of function are all central ideas in biology that are directly related to the evolutionary history of any group of organisms. Because of this, taxonomy plays an important cross-linking role in many of the NCBI tools and databases. The NCBI Taxonomy database is a curated set of names and classifications for all of the organisms that are represented in GenBank. When new sequences are submitted to GenBank, the submission is checked for new organism names, which are then classified and added to the taxonomy database. As of April 1, 2003, there were 4,653 families, 26,427 genera, 130,207 species, and 176,890 total taxa represented. Of the several different ways to build a taxonomy, our group maintains a phylogenetic taxonomy. In a phylogenetic classification scheme, the structure of the taxonomic tree approximates the evolutionary relationships among the organisms included in the classification. The NCBI Taxonomy classification represents an assimilation of information from many different sources. Much of the success of the project is attributable to the flood of new molecular data that has revolutionized our understanding of the phylogeny of many groups, especially of previously poorly understood groups such as Bacteria, Archea, and Fungi. Users should be aware that some parts of the classification are better developed than others and that the primary systematic and phylogenetic literature is the most reliable information source. NCBI Taxonomy do not rely on sequence data alone to build our classification, and we do not perform phylogenetic analysis ourselves as part of the taxonomy project. Most of the organisms in GenBank are represented by only a snippet of sequence; therefore, sequence information alone is not enough to build a robust phylogeny. The vast majority of species are not there at all, although about 50% of the birds and the mammals are represented. We therefore also rely on analyses from morphological studies; the challenge of modern systematics is to unify molecular and morphological data to elucidate the evolutionary history of life on earth.

#### Taxonomy Browser:

The Taxonomy Browser (TaxBrowser) provides a hierarchical view of the classification from any particular place in the taxonomy. This is probably the display of choice for most casual users (browsers) of the taxonomy who are interested in exploring our classification. The TaxBrowser displays only the subset of taxa from the taxonomy database that is linked to public sequence entries. About 15% of the full Taxonomy database is not displayed on the public Web pages because the names are from sequence entries that have not yet been released. TaxBrowser is updated continuously. New species will appear on a daily basis as the new names appear in sequence entries indexed during the daily release cycle of the Entrez databases. New taxa in the classification appear in TaxBrowser on an ongoing basis, as sections of the taxonomy already linked to public sequence entries are revised.

*Courtesy: National Center for Biotechnology Information, USA*

## Software Mania

### PreADMET

PreADMET is a software package for prediction of various properties based on designed structure of chemical compounds. It supports friendly user interface and MS-Windows optimized software architecture, which easily provide useful numerical information related to absorption – distribution – metabolism – excretion (ADME) and toxicity of chemical compound, from the early step of drug discovery.

#### Characteristics:

Using PreADMET for drug discovery and/or compound library design, we can perform the following tasks:

- Sketch and edit 2D structure of chemical compound & Collect structures and data fields of SD files
- Convert data into different types of file for QSA(P)R (\*.txt, \*.csv, \*.arff) or other statistical application
- Visualize 3D structures of chemical compounds
- Integrate 2D -> 3D Conversion software (Corina(Mol-net), OMEGA(OpenEye), Vconf (VeraChem), Marvin(ChemAxon))
- Calculate more than 2,000 descriptors, including both 2D and 3D descriptors
- Predict solubility in pure water or buffer solution & Screen drug candidates for similar drug group, by drug-like rule
- Predict properties related to ADME
- in vitro Caco-2 cell permeability, MDCK cell permeability & in vivo blood-brain barrier penetration, BBB
- HIA or Human intestinal absorption, in vitro skin permeability & in vitro plasma protein binding,
- Predict properties related to toxicity, Mutagenicity from Ames test (in vitro) & Carcinogenicity from mouse and rat (in vivo)
- Provide combinatorial library builder & Design chemical compound library using ADME/Tox. properties

## Bio Servers

### Popitam

Peptide identification from MS/MS data relies on a similarity measure between the experimental MS/MS spectrum and theoretical peptides from a database. In case of mutated or modified source peptides, some peaks are shifted in the spectrum, making the identification trickier. Popitam (Hernandez et al., Proteomics, 2003) is a method dedicated to the identification of peptides by tandem mass spectrometry

#### Popitam distinctive features are the following:

- Based on tag extraction, it allows identifying mutated or modified peptides, including peptides with unknown post-translational modification(s).
- The spectrum space is reduced by using database information to extract exclusively tags that are consistent with theoretical peptides.
- Several tag scenarios are built for each theoretical peptide and are scored thanks to a function which has been optimized by Genetic Programming.

**Input data parameter**

Data file :

Data file content :

Data file format :

**Other general parameters**

Database :  Release 56.7 of 20-Jan-2009  
 Release 39.7 of 20-Jan-2009  
 decoy

AC list (< 2000) :   
**Required parameter.**

Instrument :   Enzyme :    
 Fragment error :  Da  Allow up to   missed cleavage

**Scoring function (facultative parameter)**

If nothing is specified, popitam will use its default scoring functions.

Scoring file :

Scoring file content :

**Open-modification search parameters**

modGap number :

Precursor mass range  to  Da

modGap mass range  to  Da

**Output parameters**

Number of displayed peptides :

Your E-mail address

Please note that submissions are limited in time to 5 minutes.  
For longer jobs, please specify your e-mail address in the corresponding field, so that results can be sent

### Bioinfy Quiz - 014

1. The number of nucleotides on each chain turn of the DNA double helix is?

- A) 8  
B) 10  
C) 12

2. What is the main damaging effect of UV radiation on DNA?

- A) Depurination  
B) Formation of thymine dimers  
C) Single strand break

3. In C4 plants what molecule is produced from the fixation of CO<sub>2</sub>?

- A) Malate  
B) Oxaloacetate  
C) Pyruvate

4. Which of these types of molecules are precursor of prostaglandins?

- A) Alpha-tocopherol (vitamin E)  
B) C18 saturated fatty acids  
C) C20 polyunsaturated fatty acids

5. Which base modification is responsible for X chrom. inactivation in mammals?

- A) Methyl-A  
B) Methyl-C  
C) Methyl-G

## Proteomics:

### Protein Kinase Resource

The major restructuring of the Protein Kinase Resource database and Web service includes substantial changes in both data composition and user interface. The goal is to convert the Resource from a pure data-serving facility into an integrated service, which combines expanded information content and a set of tools that will allow the user to analyze the data immediately within the user interface. In addition to serving locally stored data, PKR will feature links to external resource that may contain information of interest to the researchers, such as phosphorylation sites, domain composition, chemical reactivity of ligands, etc. In order to enable these features, PKR have initiated collaborative efforts with Protein Data Bank, Alliance for Cell Signaling, Joint Center for Structural Genomics, and other online database services.

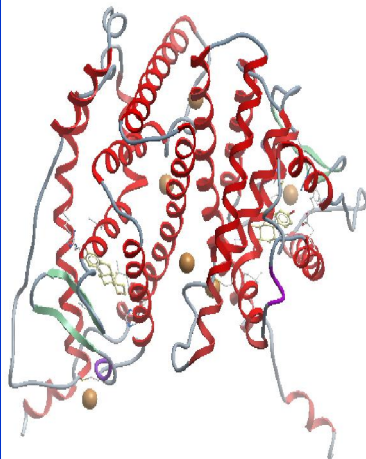
At present, PKR have established a new database structure that includes expanded table set and features comprehensive coverage of all kinase-related data and derived information, including genomic sequences, detailed organism and tissue cataloging, multiple sequence alignments, clustering into families, literature citations, and many more. In addition to sequence information, PKR are including all available structural data, which will be presented as structure alignments. This will allow for direct search and comparison of structural features of kinases of interest, including active site topology, specific features of surface contact sites and overall fold of the enzymes.

PKR is also developing a visual user interface, which will be incorporate into the PKR web pages, as well as made available as a desktop client application. The Classification Viewer is designed to display hierarchical clustering of protein kinases into classes, groups and families.

## Molecule of the Month

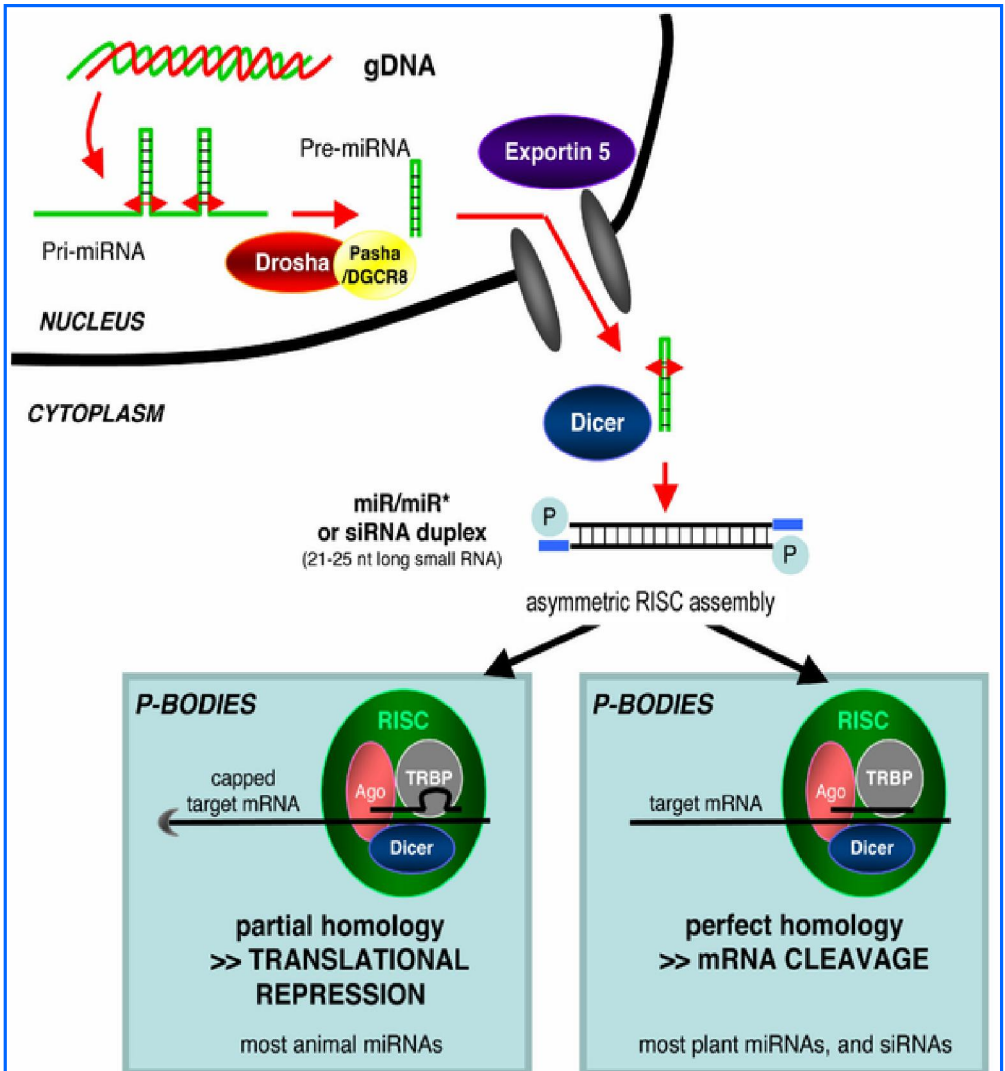
### Estrogen Receptor

Estrogens are small, carbon-rich molecules built from cholesterol. This is quite different than larger hormones, such as insulin and growth hormone, which are sensed by receptors on the cell surface. Estrogens pass directly into cells throughout the body, so the cell can use receptors that are in the nucleus, right at the site of action on the DNA. When estrogen enters the nucleus, it binds to the estrogen receptor, causing it to pair up and form a dimer. This dimer then binds to several dozen specific sites in the DNA, strategically placed next to the genes that need to be activated. Then, the DNA-bound receptor activates the DNA-reading machinery and starts the production of messenger RNA.



### Molecular Data

PDB ID : 1A52  
Amino acids : 258  
Exp. Method : X-Ray Diff  
Chains : A, B (2)



### Bioinfo Animator:- RNA Interference

Illustration of the major differences between plant and animal gene silencing. Natively expressed microRNA or exogenous small interfering RNA is processed by dicer and integrated into the RISC complex, which mediates gene silencing.

### For suggestions & contributions contact:

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

### Bioinfo Quiz

014

Answers

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