



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
Bioinfo Quiz	3
Computers for Biologist	4
Molecule of the Month	4
Bioinfo Animator	4
Contact Us	4

Adviser:

Dr. P.G. Rao

Editors:

Salam Pradeep Singh

Dr. R.L. Bezbaruah

BIF Upcoming Events

1. Training on "Application of Computer in Biological Science" @ Bioinformatics Center, Guahati University from January 21st - 23rd, 2009.

2. Training course on "Insilico Approach to Genome Analysis" @ North-East Hill University, Shillong, from February 5th - 11th 2009.

Cover Story

Biodiversity Informatics

Biodiversity Informatics is the application of informatics to recorded and yet-to-be discovered information specifically about biodiversity, and the linking of this information with genomic, geospatial and other biological and non-biological datasets. The first use of the term can be traced back as far as 1993. In 2001 Berendsohn wrote that "Biodiversity Informatics is the application of information technology (IT) tools and approaches to biodiversity information, principally at the organismic level. It thus deals with information capture, storage provision, retrieval, and analysis, focused on individual organisms, populations, and species, and their interactions. It covers information generated by the fields of systematics, evolutionary biology, population biology, and ecology, as well as more applied fields such as conservation biology and ecological management." According to Soberon and Peterson (2004) Biodiversity Informatics "includes the application of information technologies to the management, algorithmic exploration, analysis and interpretation of primary data regarding life, particularly at the species level of organization." Johnson (2007) defined Biodiversity Informatics as "an emerging field that applies information management tools to the management and analysis of species occurrence, taxonomic character, and image data." Sarkar (2007) drew an important distinction between the term "biodiversity informatics" and the term "bioinformatics", noting that "bioinformatics is an established field that has made significant advances in the development of systems and techniques to organize contemporary molecular data."

Special Interests

Tropical Botanic Garden & Research Institute (TBGRI)

An Indian Biodiversity Information Resource

Tropical Botanic Garden Research Institute, Kerala, has developed several software packages and database for the management of various types of biodiversity information. The information and resource is available on the TBGRI website: <http://www.tbgr.i.in>

Some of the database details are listed below:

Plant Info: Developed to organize a centralized digital database including all published information related to plant genetic resource and environment.

Database on Sacred groves of Kerala: List of sacred groves in Kerala with photographs, short history, vegetation types, associated flora and fauna, physiology, and other study reports of sacred grooves.

Fungal Database - Meliolales: Created for documenting the fungal diversity. This database comprises taxonomic information on fungi belonging to order Meliolales commonly known as black mildews. The database offers information on taxonomic description, Belle's formula, host details, images and reference of each species of Meliolales reported from India. It also provides general information about fungi and Indian Meliolales.

SeedPack: A software tool for the management of seed.

Computational Chemistry:

Lipinski's Rule of Five

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. 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 increased activity and selectivity, as well as drug-like properties as described by Lipinski's rule. The modification of the molecular structure often leads to drugs with higher molecular weight, more rings, more rotatable bonds, and a higher lipophilicity.

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

1. Not more than 5 hydrogen bond donors (nitrogen or oxygen atoms with one or more hydrogen atoms)
2. Not more than 10 hydrogen bond acceptors (nitrogen or oxygen atoms)
3. A molecular weight under 500 daltons
4. A partition coefficient log P less than 5

Note that all numbers are multiples of five, which is the origin of the rule's name.

Proteomics

DIP

The DIPTM (Database of Interacting Proteins) database is a catalog of experimentally determined interactions between proteins. The DIPTM database lists protein pairs that are known to interact with each other. By interact it means that two amino acid chains were experimentally identified to bind to each other. The database lists such pairs to aid those studying a particular protein-protein interaction but also those investigating entire regulatory and signaling pathways as well as those studying the organisation and complexity of the protein interaction network at the cellular level. This page serves also as an access point to a number of projects related to DIP, such as LiveDIP, The Database of Ligand-Receptor Partners (DLRP) and JDIP.

The DIP database is composed of nodes and edges:

DIP Nodes (proteins)

Each protein participating in a DIP interaction is identified by a unique identifier of the form <DIP:nnnN> and cross-references to, at least, one of the major protein databases - PIR, SWISSPROT and/or GENBANK. In addition, some basic information about each protein, such as name, function, subcellular localization and cross-references to other biological databases is stored locally (if available) in case the cross-referenced databases are not accessible.

DIP Edges (interactions)

The information about each DIP interaction is identified by a unique identifier of the form <DIP:nnnE> that provides access to information such as the region involved in the interaction, the dissociation constant and the experimental methods used to identify and characterize the interaction.

Courtesy: Swiss Institute of Bioinformatics, Geneva, Switzerland

Genomics

CAMERA

CAMERA stands for Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis. The aim of this project is to serve the needs of the microbial ecology research community by creating a rich, distinctive data repository and a bioinformatics tools resource that will address many of the unique challenges of metagenomic analysis.

CAMERA is developing the cyberinfrastructure necessary to support the data, tools and resources that will be needed to enable the scientific community to use the rapidly growing treasure of metagenomic information. Success in this will accelerate understanding of biology and deliver novel biological solutions to important societal challenges in health care, energy, and the environment.

CAMERA is making accessible raw environmental sequence data, associated metadata, pre-computed search results, and high-performance computational resources. It is based on innovative cyberinfrastructure leveraging emerging concepts in data storage, access, analysis, and synthesis not available in current gene sequence resources.

Initially, CAMERA is making available all the metagenomic data being collected by the Global Ocean Sampling (GOS) expedition, which is sampling microbial communities every 200 miles around the globe, plus 150 new full genome maps of ocean microbes. The initial incarnation of CAMERA includes two other data sets: a large-scale metagenomic survey of marine viral organisms collected from sites around the North American continent and a vertical profile of marine microbial communities collected at the Hawaii Ocean Time-Series (HOTS) station ALOHA.

Courtesy: National Center for Biotechnology Information

Software Mania

Antibase 2008



The AntiBase 2008 data collection is the ultimate tool for all scientists working in the field of natural compounds. Its author, Professor Hartmut Laatsch, began to implement his idea of a comprehensive and informative database of natural compounds in the mid-1980s. Professor Laatsch's group was the first in Germany and one of the first worldwide to investigate marine microorganisms.

Progress in natural products chemistry today means progress in synthetic chemistry tomorrow. Natural products increasingly form the basis for new medical applications. There is no doubt that the isolation and structure identification of natural products is still one of the most important areas in chemistry.

AntiBase covers over 34,400 compounds, mainly from microorganisms and higher fungi, including yeasts, ascomycetes, basidiomycetes and lichens, but also algae and cyanobacteria.

The data in AntiBase 2008 has been collected from the primary and secondary literature and then carefully checked and validated.

A demo version of AntiBase with 199 compounds and 2518 properties is available free of cost.

Bio Servers

T-Coffee

T-Coffee (Tree-based Consistency Objective Function For alignment Evaluation) is a multiple sequence alignment software using a progressive approach. It generates a library of pairwise alignments to guide the multiple sequence alignment. It can also combine multiple sequences alignments obtained previously and in the latest versions can use structural information from PDB files (3D-Coffee). It has advanced features to evaluate the quality of the alignments and some capacity of identifying occurrence of motifs (Mocca). It produces alignment in the aln format (Clustal) by default, but can also produce PIR, MSF and FASTA format. The input files are always in FASTA format.

TCOFFEE::Regular

Computes a **multiple sequence alignment** and the associated **phylogenetic tree**.
Use T-Coffee to align **Protein, RNA** and **DNA** sequences.

Limitations: Maximum number of sequences is **50**
Maximum length of sequences is **2000**

Sequence Input

switch to [TCOFFEE :: Advanced](#)

Paste or upload your set of sequences in FASTA format and Press the SUBMIT button [\[Sample File\]](#).

[Upload File](#)

FASTA Format

Browse...

or paste data

Paste a Multiple Sequence

Computation Mode

switch to [TCOFFEE :: Advanced](#)

regular: regular T-Coffee.

expresso: structural extension (E-mail recommended).

rcoffee: RNA secondary structure extension (E-mail recommended).

[Computation mode](#)

regular

You may paste your e-mail address:

Submit

Reset

Bioinfy Quiz - 007

1. In the human body what is the average rate of cycling of a molecule of ATP?

- A) 40 times a day
- B) 400 times a day
- C) 4000 times a day**

2. What is the consensus sequence for transcription factor SP1?

- A) GGGCGG**
- B) GGCCGG
- C) GGCGGG

3. Which of these proteins is not a member of the Band 4.1 family?

- A) Merlin
- B) Moesin
- C) Plakoglobin**

4. Which of these enzymes is classified as: EC 1.1.1.1?

- A) Alcohol Dehydrogenase**
- B) Carbonic anhydrase
- C) Dihydrofolate reductase

5. Sodium Dodecyl Sulphate (SDS) binds protein in which ratio:

- A) 0.8g per gram of protein
- B) 1.4g per gram of protein**
- C) 1.9g per gram of protein

Answers on Page 4

BioRuby

BioRuby is a package of Open Source Ruby code, with classes for DNA and protein sequence analysis, alignment, database parsing, and other Bioinformatics tools. Recently, tools for structural biology have been added.

BioRuby project aims to implement an integrated environment for Bioinformatics with Ruby language. The BioRuby library follows the KISS principle in order to maximize its usability and efficiency for biologists as a daily tool. The project was started in Japan and supported by University of Tokyo (Human Genome Center), Kyoto University (Bioinformatics Center) and the Open Bio Foundation.

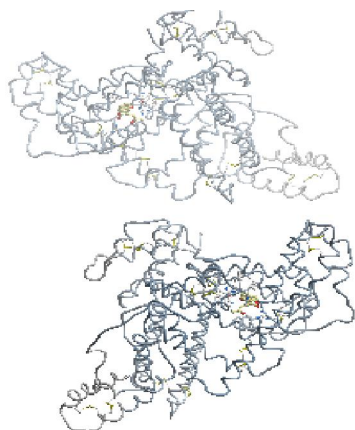
Some of the important classes available in BioRuby are:

- Sequence (translation, splicing, window search etc.)** - Bio::Sequence::NA, Bio::Sequence::AA, Bio::Location(s), Bio::Alignment, Bio::NucleicAcid, Bio::AminoAcid
- Applications (similarity search, multiple alignment, sorting etc.)** - Bio::Blast, Bio::Fasta, Bio::Hmmer, Bio::EMBOSS, Bio::Genscan, Bio::ClustalW, Bio::MAFFT, Bio::PSORT, Bio::TargetP, Bio::SOSUI, Bio::TMHMM etc.
- Data I/O (local flatfile, OBDA, KEGG API, GenomeNet DBGET, DAS, NCBI PubMed etc.)** - Bio::Registry, Bio::FlatFile, Bio::Fetch, Bio::SQL, Bio::KEGG::API, Bio::DBGET, Bio::DAS, Bio::PubMed, Bio::Fastacmd etc.
- Database parsers and entry objects** - Bio::GenBank, Bio::RefSeq, Bio::EMBL, Bio::TrEMBL, Bio::SwissProt, Bio::GFF, Bio::GO, Bio::MEDLINE, Bio::LITDB, Bio::PROSITE, Bio::DB::TRANSFAC, Bio::AAindex, Bio::KEGG::GENOME, Bio::KEGG::GENES, Bio::KEGG::KO, Bio::KEGG::ENZYME, Bio::KEGG::COMPOUND, Bio::FANTOM, Bio::PDB etc.
- Seq features, references, graphs, binary relations** - Bio::Feature(s), Bio::Reference(s), Bio::Pathway, Bio::Relation

Molecule of the Month

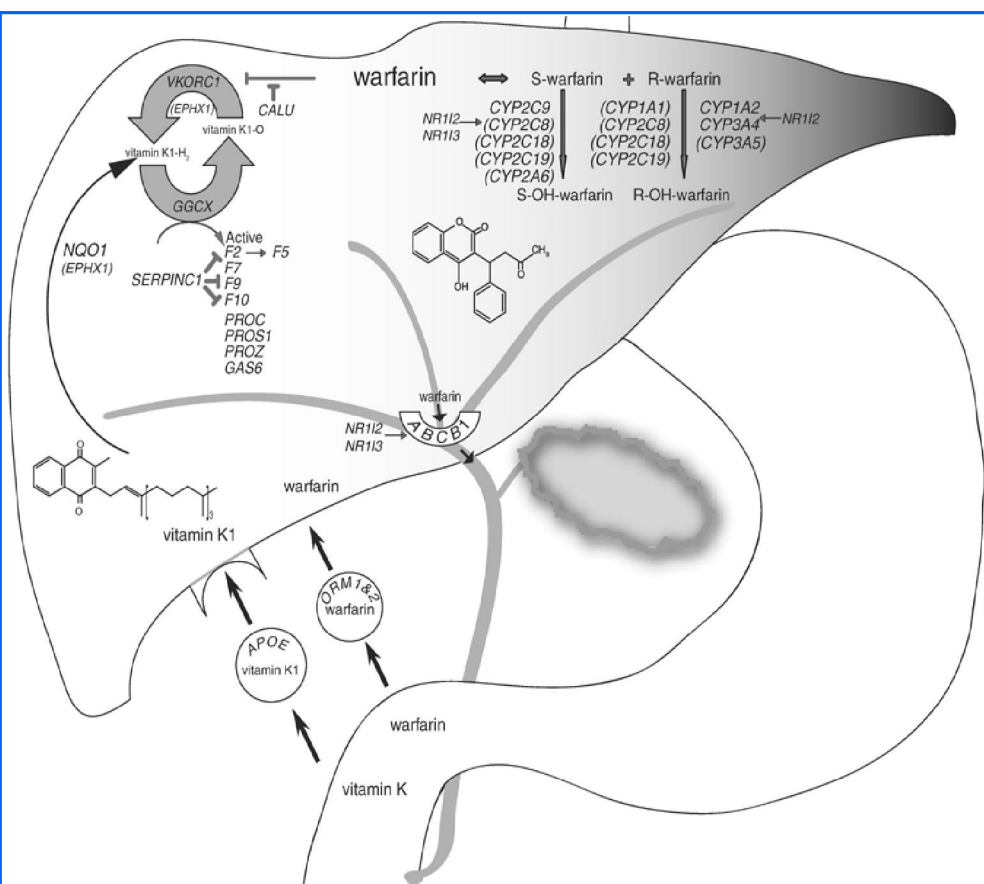
Human Serum Albumin Complexed With Warfarin

Human serum albumin (HSA) is an abundant plasma protein that binds a remarkably wide range of drugs, thereby restricting their free, active concentrations. Crystallographic analysis of 17 different complexes of HSA with a wide variety of drugs and small-molecule toxins reveals the precise architecture of the two primary drug-binding sites on the protein, identifying residues that are key determinants of binding specificity and illuminating the capacity of both pockets for flexible accommodation.



Molecular Data

PDB ID	: 2BXD
Amino acids	: 572 + 578
Atoms	: 8588
Exp. Method	: X-ray
Chains	: A & B (2)
Deposition	: 2005-07-26



Bioinfy Animator - Warfarin: Current Status and Future Challenges

Warfarin is an anticoagulant that is difficult to use because of the wide variation in dose required to achieve a therapeutic effect, and the risk of serious bleeding. Warfarin acts by interfering with the recycling of vitamin K in the liver, which leads to reduced activation of several clotting factors. The most important genes affecting the pharmacokinetic and pharmacodynamic parameters of warfarin are *CYP2C9* (cytochrome P(450) 2C9) and *VKORC1* (vitamin K epoxide reductase complex subunit 1).

Please contribute, contact:

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

Bioinfy Quiz 006
Answers

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