UNIVERSITY OF NORTH BENGAL

We're on the web!
http://nbu.edu.in/bioinformatics

Bioinfoline

VOLUME 1 ISSUE 3 SEPTEMBER 2008

INSIDE THIS ISSUE:

Editors:
Dr. Arnab Sen
Dr. Asim K Bothra

Plant Peptidomics

Life on earth is the most beautiful and delicate manifestations of ever happening interactive relationship between biotic and abiotic form, where non-livings create, connect, care and livings perform. In this regard the performance of any living organism is mostly dominated by its communication skill, the art to connect and to be connected. The execution of this skill is mainly dependent on signaling molecules; they connect cell to cell, create unity between them and even unfold the extremes. Biotic world shares lots of signaling molecules in the form of organic compounds, especially biomolecules like proteins, peptides, hormones, enzymes etc., inorganic compounds and energy. In recent years, a vast array of bioactive peptides are being isolated from different spectrum of life form and only some of these low molecular weight peptides have been characterized in details. Over the last decade it has become apparent that plants also contain peptidic signalling molecules that play vital roles in cell-to-cell communication. Plant peptides are protein molecules smaller than 10 kDa that can essentially be divided into two categories: bioactive peptides that are produced by selective action of peptidases on longer precursor proteins, and degraded peptides that result from the activity of proteolytic enzymes during protein turnover. Although both groups are products of proteolysis, they differ in how they act within the cell. The first group has key roles in various aspects of plant growth regulation through signalling, endurance against pests and pathogens by acting as toxins and elicitors, and detoxification of heavy metals by sequestration. Often these peptides bear certain sequence patterns or motifs. By contrast, the second group has no such pronounced cellular effects, but may play an important role in nutrient mobilization across cellular membranes or in function that remain to be defined.

Systemin was the first plant peptide shown to have a role in plant signalling where it functions in the systemic wound response.

(Continued on page 5)

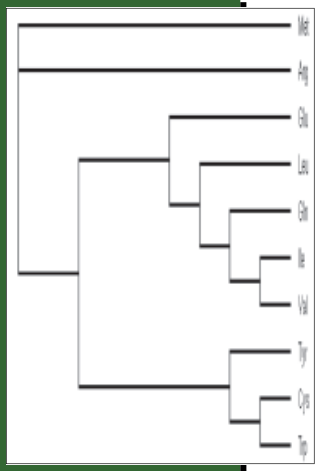
Molecular protein model validation

A protein structure determined experimentally using X-ray crystallography or NMR spectroscopy or by computational techniques like homology modelling results in a model that requires validation to establish its functional and structural integrity. While all structural models obtained contain mistakes, they become less of a problem when it is possible to detect them. It is sometimes difficult to establish a structure experimentally using X-ray crystallography or NMR spectroscopy and nowadays computational techniques have become very well-accepted in generating 3D structures. Homology modelling is a reliable technique that can consistently predict the 3D structure of a protein with precision akin to that of a protein structure determined at low-resolution by experimental means. Once an error is recognized in a model, it is feasible to categorize whether it affects the structural or functional regions. As a result a number of strategies have been developed to overcome these errors. Consequently, an essential step in the model generation process is the detection of wrongly modelled regions. Validation is an absolute necessity in the modelling procedure. Wrong models have resulted in retraction of a number of papers and structures from the public domain.

There are two dissimilar approaches to detect errors in a structure: 1) checking the consistency of the model with experimental data of the target protein, and 2) evaluating stereochemistry and other spatial characteristics of the model by means of methods based on statistics resulting from experimentally determined protein structures. Since, the explanation of the physico-chemical properties of the

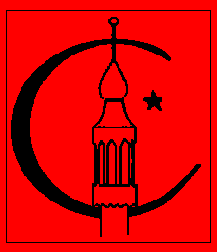
(Continued on page 6)

New method for molecular phylogeny study alternative to sequence alignment



Amino acids sequence based phylogeny has some limitation due to very low sequence similarity amongst the different tRNA synthetases and structure based phylogeny has also its limitation. In this regard, researchers at Raiganj University College, Bose Institute and BIF of North Bengal University in their study used tRNA nucleotide sequences of *E. coli* K12 (Bacteria), *Saccharomyces cerevisiae* (Eukarya), *Thermococcus kodakaraensis* KOD1, and *Archaeoglobus fulgidus* DSM 4304 (Archaea) for phylogenetic analysis. The aminoacyl-tRNA synthetases are some of the major protein components in the translation machinery. These essential proteins are found in all forms of life and are responsible for charging their cognate tRNAs with the correct amino acid. These important enzymes have been the subject of intense scientific inquiry for nearly half a century, but their complete evolutionary history has yet to emerge. Their results complement the observation with the earlier studies based on multiple sequence alignment and structural alignment. The researchers observed that relationship between archaeal tRNA synthetases are different that of bacteria and eucarya. Violation of Class rule of LysRS is observed here also. The uniqueness of this method is that it does not employ sequence alignment of complete nucleotide sequence of the corresponding gene. The paper will be published in Journal of Biomolecular Structure and Dynamics in the October issue.

BIF-NBU Wishes Everyone a Very Happy Festive Season Ahead



This newsletter is published with financial assistance from Department of Biotechnology, Gov. of India under the scheme: Promotion of Biotechnology teaching through Bioinformatics (BTBI).

Our Address:
University of North Bengal, Bioinformatics Facility, Department of Botany, Siliguri WB-734013, India
Phone: 0353-6528172, Fax: 0353-2699001
E-mail: nbengaluniv.btisnet@mail.nic.in, bif.nbu@hotmail.com

BIOINFO NEWS

Gene for deadly childhood cancer detected

Scientists have discovered the gene behind an inherited form of neuroblastoma, a cancer of the nervous system predominantly affecting children. They are sanguine that the findings will allow them to develop mechanism for disease screening in some families, as well as lead to potential new therapies.

Searching for shared DNA, the scientists promptly searched a region of chromosome 2. That led them to mutations in a gene called *ALK*, which, when activated, can promote cancer. The defective form of *ALK*, a dominant allele, appeared in all the affected individuals, as well as in healthy parents, who had passed it down. The families studied, carried the *ALK* mutation, making up a tiny minority of those affected by neuroblastoma. Mainly children are the only ones in the family carrying the disease, developing it impulsively for no obvious reason. Only a handful of children without a family history of the mutation carried *ALK* mutations in every cell in their body, implying that they developed a spontaneous mutation in utero before going on to develop neuroblastoma. Together, these facts suggested that *ALK* had a role to play in spontaneous neuroblastoma as well as the familial form, even though oncologists are still trying to find out precisely what it might be. The problem lies with the diagnosis. Most children suffering from neuroblastoma are diagnosed once the disease has already spread and have a survival rate of about 30%.

Figure: A child suffering from neuroblastoma. Thousands of children all over the world would be benefitted from this discovery.

"Mainly children are the only ones in the family carrying the neuroblastoma, developing it impulsively for no obvious reason"

UniProt released complete human proteome

UniProt consortium has released the annotated representation of all the currently known human protein-coding genes. It consists of 20,325 entries. About a third of these contain extra sequences depicting isoforms produced by alternative splicing, alternative promoter usage and alternative translation initiation, resulting in approximately 34,000 human protein sequences. About 46,000 single amino acid polymorphisms (SAPs), mostly disease-linked, have been described along with 60,000 post-translational modifications (PTMs). Its another feather in the cap of UniProtKB/Swiss-Prot. They have provided a fully annotated proteome set for a model organism (for example *E. coli* or *S.cerevisiae*) in the past and some more has been planned (*A. thaliana*, *B. subtilis*, *D. discoideum*, mouse, rice, *S. aureus*, *S. pombe*, etc). This is probably for the first time that the life sciences community has been provided with a complete set of human proteins. With the goal set to fully understand *Homo sapiens* at the molecular level this representation is expected to significantly contribute to this extraordinary adventure. Entries will be created for newly discovered human proteins, the existing sets would be reviewed and updated, the number of splice variants increased, full range of PTMs will be explored and a comprehensive view of protein variation would be built in the human population. However, the categorization at the molecular level will need to be placed in its physiological perspective i.e., subcellular location, tissue expression, protein/protein interaction, etc. This initiative is hoped to pave the way for exciting aspects. (www.uniprot.org)

BIOINFO NEWS

Archon X PRIZE for Genomics

Stephen Hawking, bestselling author of the book "A Brief History of Time" and the children's book "George's Secret Key to the Universe along with his daughter Lucy, will be sending his digitized DNA into space as part of NCsoft corporation's operation, called Operation Immortality. Together, the father and daughter are hoping the project will raise awareness of the Archon X PRIZE for Genomics, a competition that will award $10 million to the first person or team that can sequence 100 human genomes within 10 days or less. Operation Immortality is a project intended to collect and archive the very best of what humanity has accomplished by sending a digital time capsule of the human race, including messages from people around the world and DNA samples from some of our brightest minds, musicians, athletes and video game players. Hawking's DNA will be transported into space by celebrated video game developer and long-time member of the X PRIZE Foundation's Board of Trustees, Richard Garriott, who is traveling to the International Space Station (ISS) in October. Garriott, will be taking Hawking's digitized DNA as well as an electronic copy of "George's Secret Key to the Universe" on a storage device called the Immortality Drive where it will be placed on the ISS. This is not the first time Hawking and Garriott have teamed up for high-flying adventures. In 2007, Garriott hosted Hawking aboard a zero gravity flight where Hawking was able to experience a weightless environment. The "Immortality Drive," is currently being loaded with information from people all over the world at the OperationImmortality.com website. Everyone is encouraged to submit their suggestions for humanity's greatest achievements, and leave their immortalized message for future generations. A select few may also have their DNA chosen to join Garriott, Hawking and other icons on an out-of-this-world experience, and possibly become the future of mankind.
(input from: *Genetic Engineering and Biotechnology News*)

"Operation Immortality is a project intended to collect and archive the very best of what humanity has accomplished"

Genome sequence of *Trichoderma* completed

A research team of government, academic, and industry researchers led by the U.S. Department of Energy Joint Genome Institute and Los Alamos National Laboratory completed the genome sequencing of *Trichoderma reesei.* Analysis of its genome revealed a surprisingly minimal repertoire of genes that it employs to break down plant cell walls, highlighting opportunities for further improvements in enzymes customized for biofuels production. The sequencing of the *Trichoderma reesei* genome is definitely a giant step towards using renewable feedstock's for the production of fuels and chemicals. This fungus responsible for causing soft rot serves as the world's most prodigious producer of cellulases and is already a dominant source of a wide variety of cellulase products for the textile industry worldwide. It is also the organism of choice for producing enzymes for the breakdown of cellulosic biomass to fermentable sugars, which can then be biologically converted to fuels and chemical building blocks. The information contained in its genome will allow scientists to understand how this organism degrades cellulose so efficiently and to understand how it produces the required enzymes so prodigiously.

New Software for Bioinformatics

Agilent Technologies Inc., (NYSE:A) introduced Agilent GeneSpring GX 10.0, the next generation of Agilent's flagship gene expression bioinformatics platform. GeneSpring GX, considered the gold standard of desktop gene expression analysis, now offers tools for systems-level data interpretation and pathway analysis, enabling scientists to attain a new level of insight into the underlying mechanism of disease or biological process. GeneSpring GX has earned a solid reputation among gene-expression biologists, with more than 4,400 references in Google Scholar, including more than 1,600 in peer-reviewed publications. As systems-level studies become more prevalent in genomics research, GeneSpring GX 10.0 adds visualization and analysis tools for applications including alternative splicing, miRNA expression and real-time PCR. To get more insight into the underlying mechanism of disease, GeneSpring now has powerful pathway analysis capabilities. By providing a database of gene product interactions, scientists can build biological interaction networks from their genes of interest.

Peptidomics

(Continued from page 1)

Subsequently, several other peptide systems have been identified. These include ENOD40 with a role in root nodulation, phytosulfokines which function in cell division CLAVATA3 which is involved in shoot meristem organization, S-locus cysteine rich (SCR) proteins that act in self-incompatibility and RALF which arrests root growth and development. Other protein and peptide families have been identified that are antimicrobial and yet others such as POLARIS have been reported to influence plant growth. This growing number of functional peptide families has led to a reassessment of the role of peptide hormones, as well as further exploration to discover novel peptide families involved in plant signalling. Elucidating the roles of additional plant peptides is essential to understand the intercellular communication underlying plant biology. A bottleneck in the process of establishing the functions of plant peptides has been identifying the genes that encode peptides. In the era of genomics, if a gene is not annotated, it is not investigated. Due to their small size, genes encoding peptides are often missed in genome annotations. In computational gene identification, to minimize incorrectly predicting random small open reading frames (ORFs) to be genes, it is common practice to disregard ORFs below a certain size without empirical evidence of expression. This practice seeks to optimize the signal-to-noise ratio in gene finding, reflecting that the probability of spurious ORF in the in the genome occurring by chance increases as the size of the ORF decreases. Bias also is a result of the methods used to construct cDNA libraries, which typically utilize a molecular size selection step to eliminate cDNAs smaller than 400 to 500 bp.

Recently the bioinformatics approaches revolutionized the concept of finding unannotated peptides. The *Arabidopsis* secreted unannotated peptide database is now a useful tool for analyzing the diversity and functionality of signaling peptides encoded by single exon genes. One can prepare such type of database for other crop species if genome sequences are available. *Arabidopsis* chromosome sequences were searched for all ORFs encoding peptides and small proteins between 25 and 250 amino acids in length. The lower size limit is based on the idea that the average signal peptide is 22 amino acid long and the bioactive peptides in nature are as short as five amino acids long. The translated ORFs were then sequentially queried for the presence of an amino terminal cleavage signal peptide, using the neural network version of SignalP 3.0; the absence of transmembrane domains using TMHMM 2.0 and the absence of endoplasmic reticulum luminal retention sequences by eliminating proteins having KDEL or HDEL motifs at their C termini. Next, the OFRs were filtered against TAIR 6.0 for identifying ORFs outside the span of *Arabidopsis* Information Resource existing annotated genes. The putative ORFs found in this way can also be validated under *in silico* mode through identifying the hybridization intensities with at least twice the median signal values of each chip and scanning the matching coefficient of the probe sequence data taken from root, leaf, flower and suspension datasets within ORF sequences. For identifying the gene families of putative peptides, single linkage clustering can be performed using BLASTCLUST with the predicted preproproteins. The ortholog of these ORFs can also be measured by tBLASTn and BLASTp conducted for predicting the functionality of the translated products of these ORFs.

One of the serious limitations to this approach however, is that bioactive peptides are sometimes generated by limited proteolysis of larger precursors. At this time, one can neither predict which proteins encoded by plants will undergo this processing nor what peptide products would be. Also it is difficult to identify peptides lacking signal sequences which may be operated through non-classical secretary pathway. Regardless of all these limitations, plant peptidomic approaches offer novel concepts for identifying new ligands in understanding the paradigm of developmental plant biology.

ClustalW multiple sequence alignment of RALFL 6

```
RALFL6                  MAAHKKSHIR----------IFFVSVMIILSLFSGFG--EGQTYINYNGM
ath_mu_ch1_20831bottom  MAAHKMSLTS----------LFFVSIVIVLSLFSGFG--EGR-YIKYRAI
ath_mu_ch1_21704top     MQIHIFSKIKNINLIIMEARHMLVTILLLSFVFMNIMKVEAQKVIGYPAI
                        *  *  *              ::*:::::  :* .:   *.:  * * .:

RALFL6                  KGDIIPGCSSKNPKECVKIPAYSYNR--GCEISTRCQRQQHSSSS-----
ath_mu_ch1_20831bottom  AKDRVPDCT-QDPKNCVRVPVNQYHLPPGCQNTTHCYREKYHI-------
ath_mu_ch1_21704top     GRDGARGCSPKDPS-CPQQPEKPYKR--GCEKITRCERDRRKQAHLRNPR
                          *   .*: ::*. * : *   *:   **:  *:* *::

RALFL6                  ----------------
ath_mu_ch1_20831bottom  ----------------
ath_mu_ch1_21704top     KVLDVVAVMAKAKQLY
```

Lease, K. A., et al. Plant Physiol. 2006; 142:831-838

"Recently the bioinformatics approaches revolutionize the concept of finding unannotated peptides in many organisms"

-Palash Mandal

Mr. Mandal is Lecturer in Botany at North Bengal University

Announcement

National Workshop on Bioinformatics

"Genetic diversity and molecular phylogeny"

A three day National Workshop on Bioinformatics will be held at NBU Bioinformatics Facility from 7th to 9th November, 2008. The theme of the workshop is *"Use of bioinformatics tools for the study of genetic diversity and molecular phylogeny"*. Students, teachers, research scholars from biology background are encouraged to apply.

For details please contact:
The Coordinator,
Bioinformatics Facility,
Univ. of North Bengal,
Siliguri 734013, India.
Phone: 0353-6528172,
FAX:2699001
E-mail: nbengaluniv.btisnet@mail.nic.in
bif.nbu@hotmail.com

Protein model validation

(Continued from page 1)

protein and its environment is not accurate enough. Even though, new evidences are suggestive of the fact that long molecular dynamics simulations with explicit solvent could surmount errors in comparative modeling.

In the first approach experimental data is used to resolve if particular regions of the protein are correctly modeled. Biochemical data of the most important residues regarding protein overall structure and function can be used to validate the model. They must be in close proximity in 3D space and in the correct orientation to perform their role. A reliable modeling of such residues does not always ensure a good prediction; conversely, inconsistency is an important reason for concern. The primary requisite of a model is to have a good stereo-chemical quality. There are a number of software assisting in checking the stereo-chemical qualities of a model. The most commonly used program is PROCHECK that assists in evaluating the overall quality of the structure and specify the regions that require further refinement. It is very important to keep in mind that the template structures used for constructing the models must be checked before hand for their internal consistency and reliability. Another important aspect of assessing the reliability is to check the spatial features of the proteins. Here, the packing fraction, creation of hydrophobic core, atomic solvent accessibilities, distribution of charged groups, main chain hydrogen structures and atomic distances. There are a number of software associated with the use of energy profiles of the proteins to assess its structural integrity using statistical measure. ProSA is one such tool that has a large user base and is frequently used in the refinement and validation of experimental protein structures. It is a tool that is based on the statistical analysis of a number of protein structures available in PDB. Its range of application includes recognition of errors in experimentally determined structures and molecular models. The graph displayed in the figure shows the difference between one incorrect structure (A) and one correct structure (B) determined by ProSA analysis. Majority of residues in the energy plot of the incorrect structure showed positive values compared to B. An accurate structure is expected to have all residues below the zero base line. The incorrect structure (A) and the supporting publication had to be retracted from a reputed journal few years back.

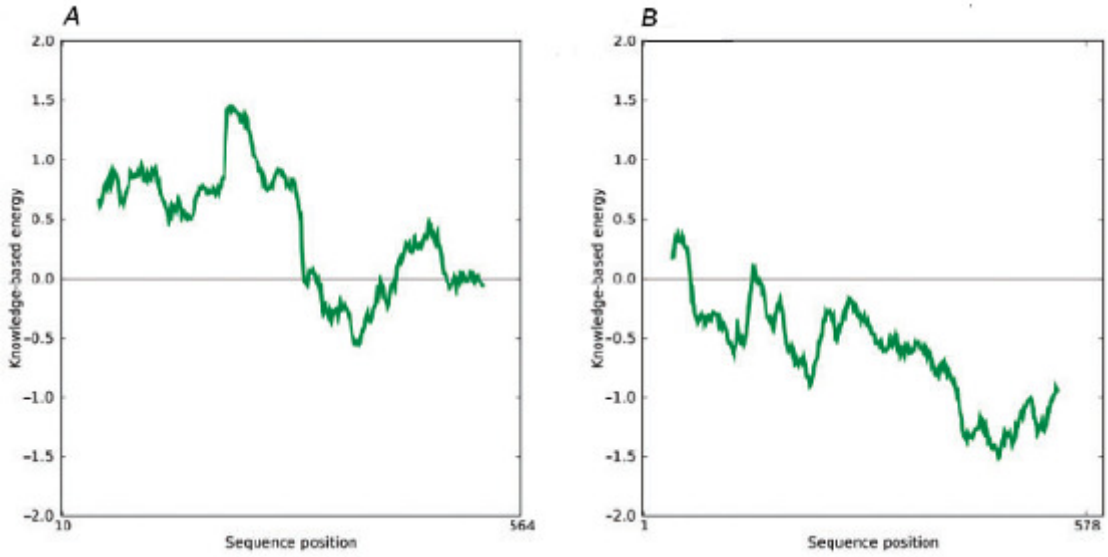


Figure: Energy plot derived from ProSA showing the residue energies over a sliding window in (A) faulty model and (B) relatively correct model. Courtesy: https://prosa.services.came.sbg.ac.at/prosa.php.

Another important tool used to validate the refined structures is VERIFY3D. Here, the 3D structures of the protein models are compared to its own amino-acid sequence taking into consideration a 3D profile calculated from the atomic coordinates of the structures of correct proteins. The constructed models of the proteins are also evaluated for their backbone conformation using Ramachandran plot. The Auto Deposition Input Tool (ADIT) (http://deposit.pdb.org/validate) can be used to examine the favorable and unfavorable qualities of the modeled structures.

During the final step of homology modeling, energy minimization and/or molecular dynamics simulations of the model can be done to decrease errors detected with PROCHECK and PROSA II. Commonly used programs for this purpose are GROMOS, CHARMM and AMBER, which explore and evaluate the multiple possible conformations of the protein.

Over and all most of the strategies focus on the suitable sampling of biologically relevant conformations of the protein useful in refining the model. This can be achieved by limiting the movement of specific amino acids.

It is highly important that these tools are applied for evaluation of the quality of models. The protein structure scientists must be aware of this fact so that grossly mis-folded structures are not deposited in the protein data bank. Most of the protein validation software are freely available in the Internet for easy access and benefit of the computational biologists.

-Saubashya Sur

Mr. Sur is Research Associate at BIF-NBU