Bifx Africa-India

Joint Virtual Conference

February 10-11, 2011

Theme: Bridging the Science between Africa and India

Organized by

Bioinformatics Organization

In collaboration with

RSG AFRICA

ASBCB

bioclues

Contents

Preface

The Bifx Africa-India Joint Virtual Conference 2011 (Bifx11) follows the success of Bifx09 and Bifx10. This joint conference is organized by the Bioinformatics Organization, and supported by Bioclues.org and African Society of Bioinformatics and Computational Biology (ASBCB).

Africa and India share some research paradigms. Over the last few years, research in the fields of bioinformatics and computational biology has gained momentum within the African continent to study interactions of pathogens, hosts and vectors in relevant diseases, even as widespread tangible research has grown within the subcontinent of India. And the "virtual conference" has taken root in these regions with the organization of several such conferences under the aegis of Bioinformatics.Org. Now, we the organizers of Bifx11 hope that this opportunity will continue to foster virtual interactions and collaborations among students, as well as researchers, of Africa and India and help to further the advancement of science there.

We have received a respectable number (14) of submissions falling under broad areas of bioinformatics and systems biology. We have six invited speakers, a tutorial, and a group discussion to sum up what will be an interesting conference for the year 2011. The Program Committee (PC) would like to thank the reviewers and sub-reviewers who have kindly reviewed the submissions. And we look forward to making the conference a success. Please find the agenda on the education website: http://www.bioinformatics.org/edu/BIFX11

-- Program Committee

Conference statistics

General Statistics

Submissions	14
Accepted	14
Acceptance rate	1.00
Reviews	27
External reviewers	5
External reviews	9

Reviewing

reviews for a paper	number of papers
1	3
2	9
3	2

Statistics by Country

country	authors	submitted	accepted	acceptance rate	PC members
Algeria	1	1.00	1.00	1.00	
Denmark					1
France	1	0.50	0.50	1.00	
India	12	6.50	6.50	1.00	1
Pakistan	13	3.00	3.00	1.00	
Peru	1	0.50	0.50	1.00	
Singapore	4	1.00	1.00	1.00	1
South Africa					2
United Kingdom	1	1.00	1.00	1.00	
United States	1	0.50	0.50	1.00	1

Invited speakers' abstracts*

* To have a glimpse of those abstracts not listed here, you may request the Program Committee

Invited speakers

 Sharmila A Bapat, PhD NCCS, Pune, India	 Simon Travers, PhD SANBI, South Africa	 Nicki Tiffin, PhD SANBI, South Africa
 Lalji Singh, PhD CCMB, Hyderabad, India	 Junaid Gamieldien, PhD SANBI, South Africa	* Jayaraman Valadi, PhD CSIR Emeritus Scientist, India

* Talk presented by Arun Gupta, CEO, AbhyudayaTech

Invited speaker

Semantic Integration of Biomedical Knowledge and Existing Data to Support In-Silico Discovery

Junaid Gamieldien, Ph D

South African National Bioinformatics Institute

Email: junaid@sanbi.ac.za

The growing number of high throughput technologies and the resulting rapid growth of biomedical data and knowledge present significant opportunities and challenges. When used in conjunction with extant biological knowledge, there exists an opportunity to re-use and re-purpose the large volumes of data in public repositories, e.g. NCBI's Gene Expression Omnibus, to generate novel biological hypotheses in-silico. Correctly integrated, such information also has the potential to facilitate disambiguation of high-throughput experimental results and to discover unobvious links between biomedical concepts through automated reasoning based approaches.

As biological information is multi-relational, often hierarchical and semantically rich, bio-integration becomes very complex when many knowledge domains are to be integrated in a way that the semantics of each and of the relationships between them remain intact. The vast majority of existing biological knowledge is also still found within biological texts and while there have been significant advances in biological text mining, strategies for integrating and utilizing this mined knowledge in conjunction with other information still have to be refined. Furthermore, as new technologies are constantly driving the discovery of new biological scenarios, relationships and relationship types, it is pertinent that a bio-integration project be developed on a database technology that is flexible enough to respond to these changes.

Relational database management systems have served bioinformatics database development well to date, but they are unfortunately not equipped to adequately deal with the semantic richness, high level of connectedness, and continuous evolution of modern biological information. Graph databases however, which focus on storing the natural structure of information in the form of a network of labeled nodes and edges (relations), enable the storage and subsequent querying of highly inter-related information. Graph databases also easily naturally represent ontologies and as long as a good semantic model is developed and adhered to, greatly simplifies the integration of complex biological information such as interspecies gene orthology, pathways, bio-entity interactions, gene-to-disease associations, relationships extracted through text mining, and even raw data/patterns from high throughput experiments. Moreover, graph databases are essentially schema-less therefore easily accommodates the addition of novel information types, which then becomes immediately accessible.

We present our integration strategy and a prototype graph database focused on human health, which seamlessly integrates hundreds of thousands of human, mouse and rat: gene, gene to disease, gene to phenotype and gene to pathway relationships and utilizes multiple bio-ontologies as 'anchors' for semantic integration. We also show how we are able to perform complex biological queries spanning multiple knowledge domains through our prototype query language.

Invited speaker

Systems Networks and Cancer

Sharmila A. Bapat, PhD

National Centre for Cell Science, NCCS Complex, Pune 411007, India
Tel.91-020-25708074, Fax.91-020-25692259, Email:sabapat@nccs.res.in

A mechanistic understanding of how genes interact in pathways, networks and complexes is critical to unravel the biological behavior of tumors towards their molecular classification. We have applied an unbiased, top-down analysis to three independent datasets of serous ovarian carcinoma, to extract a ‘SeOvCa signature’ of commonly modulated gene expressions of a prioritized list of 30 candidate genes that could have predictive or prognostic value. Signature analyses with a systems network approach affirmed importance to some genes with a pre-identified association with ovarian cancer either as predictive biomarkers, or in transformation. Our results contribute to the realization that the hallmark of serous ovarian carcinoma at the systems level is the activation of an interconnected module effected by the expression of the SeOvCa components. In the parlance of this emerging field, genes in the database displaying connectivity with these genes (designated as hubs), can be predicted to have key roles in overall network functionality. Such networking was found to involve not only genes identified by more conventional analyses in previous studies, but also added a second tier of previously unrecognized genes identified purely on the basis of expression patterns of interconnectivity with SeOvCa genes, and whose precise functions in transformation remain to be investigated.

Contributing authors:

Anagha Krishnan[2], Avinash D.Ghanate[2], Anjali P. Kusumbe[1], Rajkumar S. Kalra[1]

1. National Centre for Cell Science, NCCS Complex, Pune 411007, INDIA
2. Institute of Bioinformatics & Biotechnology, Pune University, Pune 411007, INDIA

Invited speaker

Support Vector Machines: Applications and Challenges in Bioinformatics

Jayaraman Valadi*, Scientist Emeritus-CSIR, India and APBioNet.org-India representative
and Arun Gupta$ CEO, AbhyudayaTech

$ Presenting author: arun@bioclues.org ; *Corresponding author: valadi@gmail.com

Support vector machines (SVMs) are a set of related supervised learning methods through which one could analyze data and recognize patterns, used for classification. With an emphasis on solving life problems in various fields, especially Chemo/Bio Informatics, the SVMs have been exploited in numerous applications. We discuss the various applications and challenges the SVMs pose particularly in identifying disordered regions in proteins. It has been observed that disordered regions in proteins provide conformational flexibility to the proteins. This flexibility allows the proteins to bind to antigens or other molecules like drugs. We deduce this approach by showing predictions of disorders, protein sub-nuclear localization and solubility and anti microbial peptides to name a few. At the end of the talk, we give a glimpse of Hybrid SVM –ACO for simultaneous classification and feature extraction using Weka classifier. Some of the problems that have been solved by us using the Weka and aside which allow users to easily combine them with the functionalities of Weka like classification, clustering, validation and visualization facilities on a single platform will be discussed. The bottomline is such use of classification reduces the overhead of converting data between different data formats.

Tutorial

Gene Expression data analysis and visualization

Shailender Nagpal
Bioinformatics.Org

Gene Expression is about studying the relative abundance of gene transcripts in a given cell or tissue of a particular organism. This transcript profile explains a normal or deviant phenotype through a complex network of interactions and pathways. If understood well, this can be a valuable tool in understanding the mechanism of action of diseases or drugs.

While many platforms are now able to generate high throughput measurements of transcript abundance in samples being studied, the methods for subsequent quality check and analysis are still poorly understood by biologists, who rely on bioinformaticians to take the best analysis paths. In this session, we talk about some of the technologies and algorithms/methods/workflows for analysis and visualization of expression data. Topics will include hypothesis testing, visualization plots, clustering, classification, gene enrichment and pathway analysis.

Group discussion

To encourage young scientists from Africa and India to nurture research, the PC will take opinion of participants in the form of a group discussion (GD-chat). There will be a moderator who will have a microphone to take opinions of participants. The participants will have a choice to answer and discuss by typing in the chat box of the GD room. The title of the GD is in agreement with the theme of the conference and subcontents listed is as follows:

Bridging Science between Africa and India: Grand challenges

- *Slipping backward to looking forward (10 minutes)*
- *The how of fostering competition (10 minutes)*
- *The three Cs of research: Consistency, Continuity and Creativity (10 minutes)*
- *Challenges for Africa and India: vision 2020 (10 minutes)*
- *Take home messages (10 minutes)*

Abstracts from submissions*

* The PC is not responsible for any misappropriate use of scientific content in the abstracts. The abstracts have been typographically edited by the PC.

Submission#1: ***Immunoinformatics*** ***Virtual poster***

Analysis of gp41 antigenic epitopes of HIV 1a: an *in silico* approach

Saliha Kiran[1,] Muhammad Ilyas[2], Abida Shehzadi[2] and Ayma Aftab[2]

1. Department of Bioinformatics, Government College University, Faisalabad, Pakistan
2. National centre of Excellence in Molecular Biology University of the Punjab, Lahore, Pakistan

Corresponding author: Abida_Ravian@hotmail.com

Background:
HIV infected individuals in Pakistan is about 80,000, *i.e.* 0.1% in comparison to the 38.6 million infected patients worldwide. HIV viral glycoprotein, gp41 assists the viral entry into host cell. Hence, viral gp41 of Pakistani population was analyzed for the CD4 and CD8 T cells binding epitopes as effective candidate of vaccine.

Results:
Immunoinformatics tools were applied for the study of varient region of HIV envelope protein, *i.e*. gp41. The protein nature was analyzed using freely accessible computational software. About 90 dp41 sequences of Pakistani origin were aligned and variable and conserved regions were found. Four segments were found to be conserved in gp41 viral protein. A method was developed, involving the secondary structure, surface accessibility, hydrophobicity, antigenicity and molecular docking for the prediction and location of epitopes in the viral glycoprotein. Some highly conserved CD4 and CD8 binding epitopes were also found using multiple parameters and docking analysis with HIV binding ligand (enfuvirtide) at minimum energies -302.369 KJ. The predicted epitopes mostly fall in the conserved region of 1-12; 14-22 and 25-46 and hence we believe they can used as continuous peptides as effective vaccine candidates.

Conclusion:
The study revealed potential HIV subtype a derived CTL epitopes from viral proteome of Pakistani origin. The conserved epitopes are highly useful for the diagnosis of the HIV 1 subtype a. This study will also help scientists to promote research for vaccine development against HIV 1 subtype A to save Pakistani population from potential disease.

Keywords: Gp41, HIV

Submission#2: ***Databases*** ***Virtual poster***

Pakistan’s Medicinal Plant Compounds Database+

Fizza Mughal1, Ayesha Ejaz1, Imran Manzoor2 and Muhammad Ilyas3*

1. COMSATS Institute of Information Technology (CIIT), Islamabad, Pakistan
2. Virtual University, Pakistan
3. National Centre of Excellence in Molecular Biology, University of the Punjab, Lahore, Pakistan

+Open source database

*Corresponding Author: milyaskh@hotmail.com

In order to utilize the potential of immense ethnobotanical research on Pakistan’s medicinal flora, we have developed an open source database of phytochemicals, that shall be freely available online. Being a first of its kind effort with respect to phytochemicals stemming from the botanically diverse rural areas of Pakistan, it has been constructed with special focus on integrating information of compounds. These compounds have been derived from plants used for therapeutic purposes of various skin diseases and conditions, as well as plant species possessing anti-cancer properties, based on ethnobotanical literature relevant to Pakistan’s extensive wealth of remedial flora. This database provides chemical, taxonomic and medicinal details related to the compound in conjunction with its Plant origin.

Keywords: Medicinal plants, drugs

Submission#3: ***Disease Biology and drug therapeutics*** ***Virtual poster***

Systems Biology for Drug Delivery in Infectious Diseases

Shailza Singh
National Centre for Cell Science, Pune, India
Email: shailza_iitd@yahoo.com

The reductionist approach of understanding proteins individually is obviously not sufficient, even at atomistic levels, making systems biology approaches essential to gain holistic insights. A full understanding of biological function emerges only if we are able to integrate all relevant information at multiple levels of organisation to recreate dynamic interactions. These dynamic interactions cannot be recreated purely by experimental observation and the only feasible approach is to develop mathematical and computational models which couple together underlying complex interacting non-linear processes. It is thus particularly encouraging to revisit the force field parametrization on the basis of extended QM calculations in conjunction with available experimental information. Further levels of approximation can be built with the combination of important advances in methodology and computer codes. Moreover, the error controlled strategy applied for optimizing an empirical method for phospholipids is novel in this domain. Understanding the mechanisms by which resistance arises may offer a route to addressing the insensitivity of signaling networks to drug intervention and restore the efficacy of anti-infectious therapy. Combining these strengths in an approach that builds membrane models, integrating adequate atomistic and electronic information and thus will represent a huge advance towards describing real membrane systems on a solid basis.

Keywords: systems biology, atomistic simulation, membrane

Submission#4: ***Disease Biology and drug therapeutics*** ***Virtual poster***

Solubility escalation of clotrimazole using biosurfactant based mixed micelles and its topical delivery

Gireesh Tripathi[1]* and Pankaj Jain[2]

1. Oriental College of Pharmacy, Bhopal, India
2. Sun Pharma Research Centre, Baroda, Gujrat, India

*Corresponding author: **rx.dops@yahoo.com**

Clotrimazole, which is an imidazole derivative, is widely and effectively used for the treatment of oral, cutaneous, vaginal candidiasis. Unfortunately, oral use of clotrimazole is unacceptable due to severe side effects. Plasma half-life of clotrimazole is 3-6 hours, suggesting that frequent dosing is needed. Thus topical administration of clotrimazole is recommended. However, it is limited by its very low aqueous solubility requiring it to be incorporated into a suitable vehicle. The delivery of clotrimazole by mixed micelle may help in the localized delivery of the drug; improved solubility and availability of the drug at the site may reduce the dose and systemic side effects.Clotrimazole is very less soluble in the water as well as in phosphate buffer. To enhance its solubility we proposed to give the surfactant treatment. As the surfactants are very much toxic and provide avoidable effects we select bile salt as the surfactant. Their properties are unique, *i.e.* better solubilzing agent, compatibility with many lipids, biodegradable and more acceptable. However bile salts alone it self are toxic and shows haemolytic effect. Lecithin neutralizes this effect, so that lecithin (Soya PC) is used to prepare mixed micelles. The mixed micellar system is utilized for the purpose of solubility enhancement because solubilization potential of bile salt- lecithin mixed micelles is higher than the micelles. We developed an effective mixed micelle delivery system using biocompatible surfactant by enhancing solubility of clotrimazole for superficial as well as deeper skin fungal infection.

Keywords: Clotrimazole, surfactant, mixed micelles, bile salt

Submission#5: ***Evolution and Phylogeny*** ***Virtual poster***

Bioinformatics for the analysis of hepatitis B virus in pregnant women

Silvia Vasquez[1]* and Howard Fields[2]

1. Instituto Peruano de Energía Nuclear, Peru, South America
2. CDC, United states

*Corresponding author: svasquez@ipen.gob.pe

The aim of this research was to analyze phylogenetical sequences that were 402 nucleotide long from the HBsAg gene of HBV obtained from an evaluation of pregnant Peruvian women who received health care in hospitals from Abancay, Huanta or Lima. The sequences were aligned using Clustal W and MUSCLE [1,2]. Nucleotide evolutionary distance matrices were calculated using the Jukes-Cantor and Tamura-Nei nucleotide substitution models. Amino acid distances were estimated using the pairwise distance model. Neighbor-joining bootstrapped trees were created using MEGA5 (Tamura et al. 2011) [3,4] and Phylip (ver 3.69) with reference sequences from GenBank for genotypes A, B, C, D, E, F, G and H. The trees obtained from these two methods were identical. This analysis shows that the sequences from pregnant Peruvian women were homologous with genotypes F, C and B.

Keywords: Phylogeny, HBV, pregnant, genotypes

References:

[1] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, J.D. Thompson, Nucl. Acids Res 31 (2003) 3497-3500.

[2] J.D. Thompson, D.G. Higgins, T. J. Gibson, Nucl. Acids Res 22 (1994) 4673-4680.

[3] K. Tamura, J. Dudley, M. Nei, S. Kumar, Mol. Biol.Evol. 24 (2007) 1596-1599.

[4] S. Kumar, M. Nei, J. Dudley, K. Tamura, Brief. Bioinformatics. 9 (2008) 299-306.

Submission#6: ***Evolution and Phylogeny*** ***Virtual poster***

Molecular Evolution of Placental Peptide Hormones

Syyada Samra Jafri[1]*, Zeeshan Hussain Shahid[2] ,Yasir Bhatti[3], Imran Ali[4],
Adnan Farooq Malik[5], Muhammad Ilyas[1] and Ziaur Rahman[1]

University of the Punjab, Lahore, Pakistan
Departments: 1:CEMB, 2: GC, 3: UVAS, 4: Forestry, 5: IBB

*Coresponding author: samra_syeda02pk@yahoo.com

The sequence of growth hormone is generally strongly conserved in mammals but short bursts of rapid change during evolution of primates and Artiodactyls have let to marked differences in primary structure and biological specificity in humans and ruminants GH. This study was conducted and protein sequences for Bubaline and Bovine Growth hormone for placenta and pituitary were identified. Studies were concentrated to identify the phylogeny of Bovidae family. Amino acid sequence in the new classification as a new order Cetartiodactyla has been found. Recent molecular studies have shown that Cetacea are nested within Artiodactyla and the combined grouping has been termed as Cetartiodactyla. GH, PRL and PL like proteins were isolated from pituitary and placental extract in human and local specie, Bubalus bubalis, Lahore, Pakistan. Placental extract from human and B. bubalis was processed. Placental extract from B. bubalis was further processed for purification on Sephadex G-150 by gel filtration chromatography. This gave single band on Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE) for monomer (Mr 22, 276). The bands were confirmed by Immunoblotting using antibody (anti GH antibody, raised in goat) and sequenced from St' Andrews University, Scottland, UK. These sequences were aligned, their models and Phylogenetic tree was constructed. The data accord with the idea that Cetacea should be nested within Arteodactyla and indicate that the episode of rapid growth hormone evolution in the leading to ruminants occurred after divergence of camelids and cetaceans.

Keywords: GH, PRL, PL, Cetartiodactyla, Cetacea, Artiodactyla

Submission#7: ***Machine Learning*** ***Virtual poster***

Spectral kernel clustering with point symmetry for remote homology detection

Anasua Sarkar [1]* and Ujjwal Maulik [2]

1. LaBRI, University Bordeaux 1, France
2. Jadavpur University, Kolkata, India

*Corresponding author: ashru2006@hotmail.com

Remote homology detection among proteins in an unsupervised approach from sequences is an important problem in computational biology. The existing neighborhood cluster kernel methods and Markov clustering algorithms are most efficient for homolog detection. Yet they deviate from random walks with inflation or similarity depending on hard thresholds. Our spectral clustering approach with new combined local alignment kernels more effectively exploits state-of-the-art neighborhood vectors globally. This approach combined with Markov clustering similarity after point symmetry based corrections outperforms other six cluster kernels for unsupervised remote homolog detection even in multi-domain and promiscuous proteins from Genolevures database with better biological relevance. Source code is available upon request.

Keywords: Spectral clustering, kernel matrix, point symmetry based distance measure, remote homology detection

Submission#8: ***Theoretical Biology*** ***Virtual poster***

Theoretical Studies on Plant Development

GuruDas College, Kolkata, India
Anamika Basu ashru2006@gmail.com

From the same eukaryotic unicellular, ancestor plants and animals have independently evolved multicellular development. Each kingdom separately evolved its mechanisms of development. Plants have evolved through increasing levels of complexity, from a fresh-water green alga, through bryophytes, lycopods, ferns and gymnosperms to the complex angiosperms of today. About 480 and 360 million years ago, from a simple plant body consisting of only a few cells, land plants (liverworts, hornworts, mosses and vascular plants) evolved an elaborate two phase life cycle with complex organs and tissue systems. Fundamental aspects of the plant body plan are remarkably consistent within the plant kingdom irrespective of vast diversification and are totally different from metazoans.

There are two approaches to the general theory of developmental regulation and evolution of plant kingdom. In first approach, it is considered that the evolution of developmental regulation as dictated by transcription factors and their binding sites in the regulatory regions of genes. This proposal latter extended for cell signalling molecules. There is a different point of view, where the primary evolution of developmental forms are thought to be dictated by general physical properties of cell assemblies. Advantageous forms are thought to be later stabilised by the gene regulatory networks. Examples of this process have been worked out in considerable details for the the case of the evolution of animal forms.

For animals, Carroll and colleagues have introduced the concept of a developmental genetic toolkit, i.e. morphologically different animals have some common families of genes that regulate major aspects of pattern development in body. This toolkit contains many families of transcription factors and most signaling pathways. One previous review had compared the mechanisms for cellular differentiation and cell-cell communicationfor animal and plant kingdoms and shown that since both plants and animals arose from the same eukaryotic linakage, the common toolkits (i.e. the starting sets of functional genes) are present for their development. Floyd and Bowman in 2007 proposed a similar idea, which may be applicable for plants.

It is necessary to have a detailed and comprehensive study of the developmental genetic toolkit of plants and their similarities and differences with those of animals. The availability of greater number of plant genomes and their improved annotation will considerably facilitate this study and is expected to give rise to interesting finds. Newman in 2008, 2009 proposed an even less explored area in plant development to know how these developmental genetic toolkit (comprised of developmental transcription factors (DTFs)) are associated with the concept of dynamic patterning modules (DPMs) in development. The DPMs, are almost conserved at the gene level during evolution, can function separately and in combination in developmental pattern formation. The upstream and downstream portions of DPMs are designated as the developmental transcription factors or DTFs and intracellular gene regulatory networks (GRNs) respectively. An identical set of genes are present in all cell of a multicellular organism. But in each cell type specific subsets of genes are activated

and other subsets are inactivated for specialized functions. Thus biochemical state of a cell is a dynamic one. The dynamical basis of cell type switching, depends in transcription factors and the genes that specify them. Thus DPMs are important for evolution of development in multicellular species.

In this work, we propose to carry out a very detailed and extensive theoretical study of the developmental processes in plants at different levels of resolution and compare the processes with those occurring in the animal kingdom. Since plant development is the process by which a matured plant grows from a single cell, known as zygote. Various processes are involved in development of land plant e.g. the formation of a complete embryo from a zygote; seed germination; the elaboration of a mature vegetative plant from the embryo; the formation of flowers, fruits, and seeds; and many of the plant's responses to its environment. Plant development encompasses the growth and differentiation of cells, tissues, organs, and organ systems. Plant development, as in all organisms, is basically regulated by its genetic complement, but, in contrast to multicellular animals, it is also characterized by extreme plasticity.

Keywords: Developmental biology, plant kingdom, developmental transcription factor

Submission#9: ***Software and algorithms*** ***Virtual poster***

gCut : A tool for restriction analysis.

Amir Chaouki
University of Djelfa, Algeria
Email: chaouki.amir@gmail.com

Background:
Restriction enzymes (RE) are used to digest single strands of DNA. Many restriction enzymes are used for Single strand DNA probes, intron mapping and digestion of hairpin loop standards etc. While the RE recognize and cut specific molecules along the regions, the latter area where it cuts are called restriction sites. The majority of the enzymes have sites of tetra-, penta-, hexa-, or hepta- nucleotide recognition sites. Till date, we believe there isn't an effective tool for RE analysis.

Method:
The gCut is written in python and compatible with FASTA format sequences. While it can analyze sequences of any size, the tool is portable and coherent. For circular DNA, gCut HC (for hexa enzymes) and gCut TC could be deployed for tetra enzymes whereas for linear DNA , gCut L works for both tetra and hexa enzymes.

Conclusions:
The gCut performs restriction analysis on a wide array of DNA sequences. The project is under appraisal and final review. The tool can be downloaded from http://gcut.sourceforge.net

Keywords: Restriction analysis, restriction mapping

Submission#10: ***Machine Learning*** ***Virtual poster***

Evolutionary neural networks applied to QSAR datasets in predicting IC_{50} values

Abhik Seal

Department of Bioinformatics,, Jadavpur University Campus
DOEACC Society, Kolkata 700032, India
Email:abhik1368@gmail.com

In this work, we have applied Evolutionary Neural networks (ENN) to find the best solution in terms of fitness function that selects the architecture of neural networks suited for modeling the biological activity values (IC_{50}). Normally designing neural networks[1-3] poses a problem and it consumes a lot of time for building a proper architecture that would predict the activity values closet to the experimental values, due to the design of the hidden layers and number of neurons in each of the layers. Evolutionary Neural Networks solves the problem of number of hidden layers and nodes in each layer using the operators involved, *i.e.* selection, crossover and mutation which are used in Genetic algorithm design[4]. With Machine Learning(ML) being applied to various drug discovery strategies and problems, we believe ENN could be applied in predicting the Biological activities, *i.e.* inhibitory concentration (IC_{50}) of QSAR datasets . This paper also suggests that ENNs are better predictor of biological activities other than regression methods and Neural Networks. The ENNs were applied to the whole dataset of 81 molecules which predicted R^2 of 0.94 better than multiple linear regression[5] values.The 81 molecule set have been divided in three sets of 35,44 and 66 and further, the ENNs are used accordingly making one set as train and other test, for example 35 as train set and testing on 44 as test set. The validation of the ENNs are performed by randomization of the test set. Results shows that the ENNs predicted better network architectures than normal double layered networks and Multiple linear regression.

Keywords: QSAR, Neural Networks, Multiple Linear Regression, Evolutionary Neural Networks, IC50 Correlation Coefficient

References:

[1] SS Haykin. Neural Network: A Comprehensive foundation MIT press 1994
[2] De Villers(Ed). Neural Networks in QSAR and Drug Design ,Academic Press London UK 1996
[3] Gasteiger, J and Zupan, J. Neural networks in chemistry. Angew .Chem. Int. Ed, Engl (1993) vol 32 pp 503- 527.
[4] D.E Goldberg, Genetic Algorithms in search, optimization and Machine Learning ,Addison-Wesley,New York USA, 1989
[5] Montgomery,D.C. and Peck,E.A., Introduction to linear regression analysis, John Wiley & Sons, New York 1982, pp. 34-38.

Submission#11: ***Machine Learning*** ***Virtual talk***

Hybrid Ant Colony Optimization-Support Vector Machine using Weighted Ranking for Feature Selection and Classification

Shimantika Sharma[1$] and Valadi K. Jayaraman[2*]

1. DY Patil University, Pune # Research Intern: C-DAC, Pune, India
2. Centre for Development of Advanced Computing (C-DAC), Pune University Campus, Pune 411007 MH, India

$ Presenting author. *Corresponding author: valadi@gmail.com

Background:
Most of the classification problems become difficult to be solved with the increase in the dimensionality of the datasets. These datasets may contain a large number of unimportant andredundant features that degrade the overall predictive performance. In addition, such features also add to the computational cost and have no contribution to the classification process. Feature Selection is a powerful dimensionality reduction tool that has been used in a wide range of machine learning applications. The main purpose of feature selection is the removal of non-informative, noisy and irrelevant features from a dataset, thus improving the overall predictive performance and lowering computation time without affecting the classification quality. In this work, a hybrid Ant Colony Optimization-Support Vector Machine (ACO-SVM) technique is presented which makes use of weighted ranking for efficient feature selection and classification.

Methods:
The ACO is an iterative process that is driven by heuristic information on the given feature as well as by the information about the past experiences stored in the memory of ants in the previous iterations. The heuristic information associated with a feature in this case is the weighted rank of that feature which was derived by calculating weighted sum of the Information Gain, Chi-Square and Correlation-based Feature Selection (CFS) scores of the feature. The fitness of the candidate solutions was assessed by considering the classifier accuracy of SVM algorithm that was evaluated using ten-fold cross validation. The performance of the proposed algorithm was tested using four biological datasets retrieved from the UCI Machine Learning Repository. The cross-validation accuracy was calculated for all the datasets and the results were compared.

Conclusion:
The results show that the proposed algorithm that uses Feature Selection gives better accuracy results than the algorithm which makes use of all the features within a dataset. This algorithm which is a hybrid filter-wrapper based system obtained a good subset of small number of relevant features and thus proves to be beneficial for improving the classification accuracy and increasing the simplicity of machine learning problems by removing irrelevant and noisy features within the dataset..

Acknowledgement: Dr. VKJ gratefully acknowledges financial support from CSIR, New Delhi, India for financial support in the form of emeritus scientist grant.

References:

1. Diwakar Patil, Rahul Raj, Prashant Shingade, Bhaskar Kulkarni, V K. Jayaraman, Combinatorial Chemistry & High Throughput Screening, 2009, 12, 507-513

2. M. Sadeghzadeh, M. Teshnehlab, World Academy of Science, Engineering and Technology 64 2010

3. Marco Dorigo, Vittorio Maniezzo and Alberto Colorni, IEEE Transactions on Systems, Man, and Cybernetics Part B, Vol.26, No.1, 1996, pp.1-13.

Submission@#12: ***Software and algorthms*** ***Virtual poster***

Hybrid Genetic Algorithm in Docking: A Review

Akanksha Gupta
IMS Engineering College, Ghaizaiabad, India
Email: biozen.agarwal@gmail.com

Computer-based methods for predicting the structure of ligand–protein complexes or docking algorithms have application in both drug design and the elucidation of biochemical pathways. The number of solved structures of ligand–protein complexes now permits the testing and validation of docking algorithms by comparison of predicted complexes with structures extracted from protein databases. In this paper, hybrid algorithms and their significant improvements with respect to applying only GA to docking are reviewed.

Keywords: hybrid genetic algorithm, docking, genetic algorithm, tabu search

Submission#13: ***Databases*** ***Virtual talk***

Large Scale Semantic Data Integration and Analytics through Cloud: A Case Study in Bioinformatics

Tat Thang[1], Michael Li Qing An[2], Lee Bu Sung, Francis[1], Kanagasabai Rajaraman[2]

1. School of Computer Engineering, Nanyang Technological University, Singapore
2. Institute for Infocomm research, A*Star, Singapore

Corresponding author: kanagasa@i2r.a-star.edu.sg

Advances in high-throughput sequencing has led to a deluge of biological data that is fast outpacing the capabilities of in-house computing infrastructures. The exponential growth in data increasingly demands fast, large-scale and cost-effective computing strategies to facilitate useful and advanced bioinformatics solutions. On another direction, a major part of the information to support biomedical knowledge discovery is available on large number of heterogeneous databases in both structured and semi/un-structured formats, and often the data needs to be integrated for analysis and interpretation of experiments. For example, "omics" studies often require integrating experimental data with several sequence annotation tracks reported in different data bases (Entrez/NCBI, UCSC Genome browser, GeneCard, different Affimetrix ChIP platforms, etc). Traditionally data integration is done using the warehousing approach wherein data from different sources is translated into a monolithic warehouse and the queries are executed on this warehouse. However, this poses maintenance challenges if the source databases evolve or new data needs to be integrated later on. Furthermore, much of the digital information is available as semi-structured (e.g. XML) and unstructured data, and little of this makes its way into data warehouses. Specifically, in the biomedical domain, published full text articles are arguably the most important source of information as these contain peer-reviewed high-quality facts and conclusions [Jensen *et.al.*2006]. Unfortunately these are almost completely unstructured and written in expressive prose with frequent use of domain terminology and abbreviations. Towards addressing these challenges, we propose a novel cloud-based biological data analytics framework that brings together ideas from cloud computing, semantic technologies and integrative data mining.

The core of our framework comprises two modules namely, the Data Infrastructure Module and the Data Analytics module. The Data Infrastructure Module focuses on efficient data management and computational resource management in the cloud. This is to support the data analytics services running in the upper layer. The data analytics module is responsible for providing the next generation data analytics experience to users of analytics. This module helps users to execute their initial analytics requests, but at the same time fetches useful services from the cloud, packages and presents them to users as both recommendation and information to provoke the users to formulate better analytics problems and approaches. Both the modules are implemented on top of the Knowle platform - a generic, scalable semantic technology platform that brings together a combination of semantic based technologies; text mining, ontology population and knowledge representation, in the construction of a knowledgebase upon which are deployed data mining algorithms and visual query functionality (http://semantics.i2r.a-star.edu.sg). Integrated together, the modules will invoke the relevant methods and submit tasks to the Hadoop MapReduce framework for execution. The Hadoop framework will extract the post-processed data from the underlying storage system and performs the necessary computation. The output is then returned back to modules for subsequent processing, and the workflow continues.

As proof-of-concept, we have implemented1 the core of the cloud-based analytics workflow on a biological data mining scenario that involves crawling UniprotKB and PubMed, integrating the retrieved data and then performing cross-querying across them. PubMed comprises over 20 million biomedical abstracts and UniProtKB contains over 13.5 million semi-structured records, and so integration and cross-querying of these two hetergeneous data sources alone represents a significant challenge with existing computing infrastructures. Our implementation includes an user interface to get input specs and subsequently invoking Hadoop MapReduce framework to crawl and aggregate data from Uniprot and PubMed. We have scripted an OWL-DL ontology to semantically integrate the two data sources using the Knowle platform that performs schema extraction, literature mining, etc. After suitable reformatting the integrated data is deposited into HDFS storage. We have built simple keyword search and also visual query interfaces using the ontology as query model, to retrieve specific segments of data, and conclude that the proposed cloud framework offers a promising solution for large scale semantics-based data integration and analytics.

Acknowledgments: Special thanks to Sundarrajan VS, NTU, who provided biological domain expertise and key assistance in sourcing the data via Hadoop.

Submission#14: ***Databases*** ***Virtual poster***

TRYPANOCYC – A METABOLIC PATHWAY DATABASE FOR *Trypanosoma brucei*

Bridget Chukualim
The Gambia (ITC), United Kingdom
Email: bchukualim@yahoo.com

Trypanosoma brucei is a single celled parasite, subspecies of which are causative agents of African sleeping sickness in humans and nagana in domestic animals. Sleeping sickness is endemic in about 36 countries in Sub-Saharan Africa and is threatening over 60 million people. Disease prevention is a priority, as few effective drugs are available. Currently, there are only four drugs used as treatment: Pentamidine, Suramin. Melarsoprol and Eflornithine utilizing the following targets, S-adenosyl methionine decarboxylase, dihydrofolate reductase, thymidine kinase, glycerol – 3- phosphate dehydrogenase and trypanothione. They have poor safety profiles, challenging treatment regimes and lack in efficacy. There is therefore a great need to identify novel therapeutic agents for the treatment of sleeping sickness and fortunately, efforts have intensified over recent years. One such development has been the availability of the Pathway Tools Software. We have used this production software to set up TrypanoCyc, a metabolic pathway database for T. brucei [http://biocyc.org/TRYPANO/]with the added advantage of computational identification of potential drug targets.

Keywords: TRYPANOCYC, Biopterin, Trypanothione

Submission#15: ***Annotation and Curation*** ***Virtual poster***

Computational prediction of possible metabolic route in *Phanerochaete chrysosporium* using proteome separated by 2D-Gel Electrophoresis

Sivaramaiah Nallapeta[1]*, Ranganath Gudimella[2], Shivani Chandra[1]
and M krishna Mohan[1]

1. Birla Institute of Scientific Research, Statue circle, Jaipur-302001 India.
2. CEBAR,University of Malaya, Kaulalumpur,Malaysia

*Corresponding author: nallapeta@bioclues.org

The complete genome sequence of *Phanerochaete chrysosporium* has provided essential genome data enhancing the understanding of lignin biodegradation, that plays a pivotal process in the global carbon cycle. However, datasets dealing with specific genes, pathways are not available. Hence in our study curation of raw sequences to predict the proteins and possible metabolic route under ligninolysis using various bioinformatics tools was attempted. Protein expression pattern under ligninolytic conditions using 2DE was performed and putative identity of separated proteins was assigned using Gene Ontology (GO) annotation tools. This study revealed computational methods that can be applied to predict possible metabolic routes in *Phanerochaete chrysosporium* during ligninolysis thus unraveling their possible functional interactions.

Keywords: *Phanerochaete chrysosporium*, Gene Ontology, Ligninolysis

Program Committee

Africa:

- Nelson Ndegwa, Technical Co Chair
- Kavisha Ramdayal, Chair
- Stanley Mbandi-Kimbung
- Vijayaraghava S, PhD

India:

- Pritish Varadwaj, PhD

Bioinformatics.Org:

- J.W. Bizzaro
- Prashanth Suravajhala

Sub-reviewers*:

- Tiratha Raj Singh, PhD
- Alfredo Benso, PhD
- Erich Grotewold , PhD
- Pandjassarame Kangueane, PhD
- DPS Verma, PhD
- Venkata Satagopa
- Louxin Zhang, PhD

* List of sub-reviewers not including the speakers. The speakers have also been asked to judge the submissions.