Kolkata International School cum Conference on Systems Biology

KOLSYSBIO

December 29'th 2012 to January 3'rd 2013

Venue: Auditorium, Saha Institute of Nuclear Physics, Kolkata

Abstract Book

Saha Institute of Nuclear Physics
Kolkata, INDIA

Abstracts of School Talks

KOLKATA INTERNATIONAL SCHOOL CUM CONFERENCE ON SYSTEMS BIOLOGY (KOLSYSBIO)

December 29, 2012 to December 31, 2012

CONVENORS:

Pradeep K Mohanty, SINP
Soumen Roy, Bose Institute

S01: Introduction to molecular biology: systems (modularity) all the way down

James Shapiro
The University of Chicago, USA

The lecture will introduce the students to some classical examples of molecular cell biology. The emphasis will be on how the cell functions as a system through molecular interactions. Another main point is that each of the interacting molecules itself functions as a multi-component system.

S02/S04: Structural bioinformatics: applications of tools and principles

Prasanna Venkatraman
Advanced Centre for Treatment, Research and Education in Cancer, Mumbai

The tutorial will include an introduction to structural bioinformatics from a biologist's persepctive, importance of protein-protein interactions, rules that govern these interactions, their importance in biological function.

Then there will be a discussion on bioinformatic studies that dwell on these protein-protein interactions.

Subsequently our own thoughts and the use of these prinicples in our research which will be illustrated using various examples.

S03: Introduction to quantitative analysis for biologists

Srinivasan Ramachandran
Functional Genomics Unit,
CSIR-Institute of Genomics and Integrative Biology, Delhi

These times are the age of big data. The availability of various types of big data is changing the approach to analyze and develop new strategies for drug targets, drugs, vaccine candidates to combat infectious diseases. We present here some approaches to analyze examples of big data: screening of chemical molecules for identifying potential inhibitors, high scoring immune epitopes, single nucleotide variations between strains of pathogen.

ParallelVSR: a cheminformatics pipeline for parallel virtual screening on R platform

We are recently receiving access to small molecule databases in the public domain. The chemical search space has therefore widened. New strategies need to be developed for efficient virtual screening (VS). In recent times we also are gifted to receive many screening algorithms these includes Vina, ChemmineR, etc. Computer speed has now become the technical issue and this requirement needs to be addressed without compromising accuracy. Such requirements are now being met through the development of freely available, fast and easy to use virtual screening pipeline, which is based on integrated approach, has potential for use in the field of drug discovery.

We describe here ParallelVSR a cheminformatics pipeline in R platform. We have made efforts to in integrate sequentially the ligand based virtual screening (LBVS) and structure based virtual screening (SBVS). LBVS used when the three dimensional structure of target protein is not available but one or several of its inhibitors are known and in that case structurally similar molecules to the known ones are sought from a chemical library of large size assuming that these structurally similar molecules may have similar activity. SBVS requires the knowledge of three-dimensional structure of the target protein. In this method, a set of small molecules are docked into the protein, and the stability of the protein-ligand complex and the details of ligand interaction with the active site of the target protein are the measures of activity of the docked ligand.

Starting with a small molecule as an input query the pipeline system performs fast LBVS using ChemmineR. Through this we can reduce a large chemical

library of small molecules to a manageable size and then followed through relatively slow SBVS using AutoDock Vina to rank the screened molecules on the basis of docking energy. We used data level parallelization for both LBVS and SBVS using Message Passing Interface (MPI) as communication protocol. We developed ParallelVSR comes with easy to run R *scripts*, which can run on MPI enabled system. This development offers added advantage of simplicity in using this pipeline, Parallel VSR is designed to meet the challenge of speed in VS.

Immunological Data Modeling, Scripting and Analysis of Pathogen Genomes

Genome sequence data from various pathogenic species and strains open new opportunities for development of new therapeutics and vaccine candidates through Bioinformatics analysis. Analysis of data is now possible using structured scripting. R is a programming language integrated with an R environment, facilitating easy and rapid data analysis with the help of its integrated suite of software facilities.

In the genomics era, vaccine targets prediction starts by using bioinformatics analysis of microbial genome sequences. This approach is called Reverse Vaccinology (RV). As newer bioinformatics algorithms develop and appear in public domain, the original approach can now be complemented by enhanced RV, which uses additional algorithms for prioritizing the vaccine candidates. We start with collection of known vaccine candidates and a set of predicted vaccine candidates from whole genome sequences of the selected pathogen genomes. These predicted vaccine candidates are adhesins and adhesin-like proteins identified by using various algorithms like MAAP for plasmodium species, FungalRV adhesin predictor for human pathogenic fungi and SPAAN for bacterial species. These sequences are analyzed through publicly available algorithms to obtain additional information on Ortholog, Paralog, BetaWrap Motifs, Transmembrane Domains, Signal Peptides, Conserved Domains, Similarity to human proteins from Human Reference Proteins, T-cell epitopes, B-cell epitopes, Discotopes and Allergens prediction. This combined approach forms the basis for enhanced Reverse Vaccinology. The combined data produced through these integrated analyses can be interogatted through well structured decision trees. We prepare scripts running in object oriented mode. From these results we derive a set of most probable adhesin vaccine candidates with additional qualification. Furthermore, the degree of conservation of the epitopes can also be investigated. These results could enable the development of epitope based vaccines in future.

S05: Case examples of cheminformatics, integrated immunoinformatics and single nucleotide variations

Srinivasan Ramachandran
Functional Genomics Unit,
Institute of Genomics and Integrative Biology, Delhi

Analyzing single nucleotide variation in Pathogenic Bacteria
Single nucleotide variations (SNVs) are variations in the nucleotide at single positions. Single nucleotide variations in the genome can have varying consequences depending on its location in the genme. We developed a pipeline of scripts integrating multiple algorithms through R platform. Here we show one case example for multiple strains of *Mycobacterium tuberculosis* of CAS spoligotype from India. In the present analysis, genome sequences of three strains of *M. tuberculosis*: AHHX, AHHY and AHHZ were analyzed for the single nucleotide variations with reference to the H37Rv, a pathogenic laboratory strain. Total SNVs were 1310, 1286 and 1131 in AHHX, AHHY and AHHZ respectively with reference to H37Rv. These SNVs covered 0.030%, 0.029% and 0.026% of the reference genome, in AHHX, AHHY and AHHZ respectively. Since the protein coding percentage of H37Rv is 91.3%, the SNVs were found more in the coding region than in intergenic region. Transition to transversion ratios (Ts/Tv ratio) were also calculated in each strain and analysis of SNVs supports the transitional bias in all the strains. Further study on the synonymous and non-synonymous SNVs in different functional categories of genes as per classification given by TubercuList suggests that 10-11% in cell wall and cell processes, 7-8% in conserved hypotheticals, 10-11% in information pathways, 7-9% insertion seqs and phages, 9-10% in intermediary metabolism and respiration, 13-15% in lipid metabolism, 13-17% in PE/PPE, 9-10% in regulatory proteins,7-8% in virulence, detoxification, adaptation, 6-7% in unknown and 2% of the total SNVs were not categorized. The length truncation of open reading frames due to SNVs was also detected in the proteins produced by the corresponding coding transcript. The number of SNVs causing decrease in length was 10 in each of the three strains whereas SNVs causing increase in the length of AHHX, AHHY and AHHZ were found to be 8, 8 and 7 respectively. Our findings show that majority of truncation in length occur due to substitution of G/C–>A/T whereas increase in length occur due to substitution of A/T–>G/C. Since the GC content of H37Rv is 65.9%, these single nucleotide variations might have role in increase or decrease of the GC content in the transcripts of the Indian strains of *M. tuberculosis*.

S07/S11: Synthetic genetic engineering

Calin Guet
Institute of Science and Technology, Vienna, Austria

Synthetic biology engineering techniques are often used in systems biology experimental approaches. I will discuss the potential and constraints of such engineering approaches. Emphasis will be placed on the power that synthetic biology has to address fundamental questions of living matter and its limitations in engineering biological machines.

S08: Modelling intrinsic and extrinsic noise in biochemical networks

Vahid Shahrezaei
Department of Mathematics, Imperial College, London

Gene expression is significantly stochastic making modeling of genetic and biochemical networks challenging. This stochasticity arises because of both inherent stochasticity in biochemistry (intrinsic fluctuations), as well as interactions of the system of interest with other stochastic systems in the cell or its environment (extrinsic fluctuations). In this talk I describe different approaches used to model such stochasticity.

S09: DNA-based genetics teaches us how cells control genome change

James Shapiro
The University of Chicago, USA

The lecture will introduce the students to the molecular basis of mutation and genome change. The main point will be that ALL genome changes result from dedicated biochemical activity. These activities are subject to cell regulatory circuits. Thus, the cell has the ability to change its genome when necessary.

S12: Rule-based modelling in systems biology

Vahid Shahrezaei
Department of Mathematics, Imperial College, London

Much of the complexity of biochemical networks comes from the information-processing abilities of protein modifications and, and binding to the receptors, ion-channels, signalling molecules or transcription factors. The large number of modifications and binding events produces a combinatorial increase in the size of such reaction networks. Rule-based modelling is a framework that is devised to deal with the combinatorial complexity. An additional problem is combinatorial increase in the number of parameters required to fit experimental data as the number of protein interactions increases. It therefore challenges the creation, updating, and re-use of biochemical models. I discuss a rule-based modelling framework that exploits the intrinsic modularity of protein structure to address regulatory complexity. This methodology provides a basis for scalable, modular and executable modelling of biochemical networks in systems and synthetic biology.