Genomic Med. (2008) 2:217–240
DOI 10.1007/s11568-009-9081-x

ABSTRACTS

Bioinformatics and statistical genomics

001: A physical map of human Alu repeats cleavage by restriction endonucleases

Murat A. Abdurashitov, Victor N. Tomilov, Valery A. Chernukhin, Sergey Kh Degtyarev

SibEnzyme Ltd., 2/12 Ak. Timakov Str., Novosibirsk 630117, Russia

The Alu family of DNA repeats, which belongs to SINE group, is one of the most abundant and well characterized repetitive elements in human genome. The total number of annotated Alu sequences is more than 1 million copies and their fraction in genome is about 10%. Earlier we have shown that visible bands on electrophoretical gels after total human DNA digestion are formed mainly due to cleavage of Alu repeats. We have proposed a method of restriction analysis of Alu repeats sequences in silico using specially designed software and diagram plotting. Comparison of theoretical results to those obtained experimentally has shown a good correlation in number and intensity of DNA fragments produced by total human DNA cleavage with different restriction enzymes. Thus, the obtained data allowed to draw maps of the whole set of human Alu repeats for these restriction enzymes. The suggested method of Alu repeats analysis allows to simplify a study of human DNA digestion with restriction endonucleases considering set of Alu repeats sequences instead of the whole human genome. Thus, time and efforts, which are necessary for such calculations, may be significantly reduced.

002: Prediction of deleterious human membrane transporter polymorphisms

Vishal Acharya, H. A. Nagarajaram

Center for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad, India

Human membrane transporters play direct roles in the absorption, disruption and elimination of nutrients, ions and many drugs. Human membrane transporters have been implicated in genetic disorders caused by transporter malfunction such as glucose malabsorption and insulin-resistant glucose transport. It is said that transporters play a second important role in pharmacology in that about 30% of the commonly used prescription drugs target transporters. Variations in membrane transporters contribute to inter-individual differences in the responses to various drugs. Variations are mostly due to non-synonymous single nucleotide polymorphisms (nsSNPs) some of which have been associated to certain diseases. However, large number of nsSNPs have not been associated to any of the known diseases and have remained uncharacterized. In addition to these, on-going genetic studies and human population based studies have also been leading to the discovery of new SNPs. It is important to predict the effect of nsSNPs at the molecular level as well as at the physiological level. In this study we have focused on the nsSNPs which occur on the transmembrane part of the human membrane transport proteins and try to develop a tool which can classify uncharacterized nsSNPs into benign and disease nsSNPs. In order to develop such a tool we have investigated and discovered a number of sequence-based properties and features such as burial status, evolutionary characteristics etc., at SNP sites as well as their flanking regions and used those features to develop a SVM based classification method. In this presentation we report the details of our studies as well as the results obtained.

003: ISSPred: identification of Intein proteins and their splice sites

Hifzur Rahman Ansari, G. P. S. Raghava

Bioinformatics Center, Institute of Microbial Technology, Sector 39-A, Chandigarh, India

Protein Post-translational Modification (PTM) is a common phenomenon in biology which regulates the function of proteins. Protein Splicing is a unique PTM in that it leads to cleavage of protein into internal (intein domain) and flanking (extein domain) fragments. Extein sequences later ligate together to form fully functional active protein. The process of precise intein splicing and formation of specific peptide bonds has been exploited by researchers to develop many novel applications in the field of protein engineering, enzymology, microarray production and target detection etc. Thus it is important to develop tools for the identification and characterization of Inteins. In this study, attempts have been made to predict intein proteins, domains, and their sites. Intein data obtained from InBase Intein database. In order to predict Intein proteins, we analyzed amino acid composition of intein proteins/domain and observed preference for certain type of residues. Support Vector Machine (SVM) models have been developed for predicting intein proteins using amino acid and dipeptide composition and achieved maximum MCC 0.63 and 0.77, respectively. Secondly SVM

models have been developed for predicting intein domains in protein using amino acid and dipeptide composition and achieved maximum MCC 0.76 and 0.87, respectively. Finally SVM models were developed for predicting splice sites using different window length and achieved maximum MCC 0.86 and 0.93 for N-splice and C-splice sites, respectively. This study is the first attempt to predict intein proteins, domains and their splice sites. Based on above models a prediction server ISSPred has been developed by using CGI-Perl and HTML. The Web server is available at http://www.imtech.res.in/raghava/isspred/.

004: Unleashing the unstructured proteins as potential drug target: a case study on *Mycobacterium Tuberculosis*

Meenakshi Anurag, Namit Bharija, Saurabh V. Laddha, Debasis Dash

G.N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, CSIR, 254, Okhla phase III, New Delhi, India

Intrinsically unstructured proteins (IUPs) occur across all kingdoms of life. Such proteins are capable of acquiring a dynamic ensemble of structures, in contrast to their structural counterparts. With very little attained so far with structure based rational drug discovery approaches, disrupting protein-protein interactions involving disordered proteins is being considered vital for drug targeting. Our study on *Mycobacterium tuberculosis* H37Rv strain reveals that seven percent of the proteome comprises of IUPs. Studying further for essentiality, homology with known structures in PDB and lack of homology with human proteome led us to 41 proteins. On the basis of druggability, coupled with structural bioinformatics, pathway analyses and protein-protein interaction studies, we identified a set of proteins that include PknG, CtpH, RocA, FtsW, AroF, RpoB. Tracking down to PknG—Serine/Threonine Kinase and AroF—a probable chorismate synthase that has been labeled as attractive drug targets against the 'Neglected Disease', ascertains our methodology. Our other remarkable finding is a FtsW-like membrane protein belonging to SEDS family, stabilizes the cytokinetic ring of FtsZ—another highly studied target. FtsW is a late recruit in the divisome assembly pathway and hires Penicillin-binding protein (PbpB) for the peptidoglycan synthesis can be foreseen as crucial target to crack the pathogen. A detailed insight of the structural, functional and interactive aspect of these proteins is being studied that would enable us delineate these as potential drug targets.

005: Beyond next gen sequencing: applying bioinformatics to obtain biological answers

Arif Anwar

Synamatix Sdn Bhd, 27-9, Level 9, Signature Office, Midvalley City, 59200 Kuala Lumpur, Malaysia

A key application of high-throughput low cost next generation DNA sequencing technologies is in the fields of personalised healthcare, biotechnology and agriculture.

Large scale sequencing, e.g., human transcriptomes and genomes, generates very large amounts of sequence data that need to be rapidly mapped, analysed and mined.

By using applications built upon the SynaBASE structured network database platform, data can be analysed faster than it can be produced by next generation sequencers.

A variety of applications have been developed to handle challenging data characteristics presented by these new sequencing platforms, such as short read length, mutations resulting from artefacts, 3′ sequence quality degradation and base-calling errors. Several algorithms, and statistical and software-based methods are used to identify and verify mutations in data generated from next generation sequencers.

Mutations such as SNPs, insertions, deletions, copy-number changes, structural variations and chimeras can be ranked, calculated and validated.

For example, chimeras generated during sample preparation prior to sequencing, can be identified and filtered by setting rules to remove them. It was discovered that potential chimeras can be found in 1–2% of the reads. Out of these, however, true chimeras number < 1 in 100.

Finally, inversions, deletions and insertion mutations can be identified by analysing paired-end read orientation and insert size distribution.

006: Using clusters of short invariant peptides in Pep cluster to explore protein function

Aparna Basu, S. K. Brahmachari

Institute of Genomics and Integrative Biology (CSIR), Mall Road, Delhi 100007, India

Rapid sequencing of new proteins has generated a need for new methods for functional annotation that link protein sequence directly to function, as structure determination is computationally and experimentally cost-intensive. Proteins performing the same function in different organisms vary considerably in their sequence. It may be assumed that at least structures relevant to protein function are preserved through evolution. Inverting the hypothesis, it can be argued that invariant sections on a protein sequence could act as 'signatures' for predicting the function of a new protein. A good signature should uniquely map to proteins of a given function. Multiple signatures for a single protein would increase the confidence level of function prediction except when signatures map onto the same invariant region of a protein, suggesting redundancy. For completeness, the library should have signatures corresponding to all possible protein functions.

The PepCluster database addresses this issue by creating a library of short sequence 'signatures' from an earlier library, CoPS, built on the principle of invariant peptides (Prakash et al. 2004). The 12,076 signatures in CoPS corresponded to 1250 different functions, or an average of about ten signatures per function. In practice this number can be much larger for some functions, as the signatures are not distributed uniformly over the functions. The high ratio of signatures to function, together with numerous cases of overlapping signatures, suggests redundancy in CoPS. To reduce redundancy, similar signatures have been clustered based on sequence similarity to obtain a set of 5620 consensus signatures. These signatures are used in the PepCluster web tool for prediction of function of unannotated proteins (Basu et al. in preparation; Taneja et al. poster in this conference).

In this paper, we take a detailed look at some individual clusters, to explore sequence variations that leave function unchanged, as well as those that lead to evolution of function.

References

Prakash T, Khandelwal M, Dasgupta D, Dash D, and Brahmachari SK (2004) CoPS: comprehensive peptide signature database, bioinformatics 20(16):2886–2888, http://cops.igib.res.in/copsv2/about cops.html

Basu A, Suchir, Brahmachari SK, PepCluster: I—a minimal set of conserved peptide signatures for protein function prediction with high accuracy (in preparation)

Taneja B, Suchir, Basu A, Brahmachari SK, PepCluster: a webtool to annotate bacterial proteins by consensus and core invariant peptide signatures (poster in this conference)

007: Evolution of the mammalian transcription factor binding repertoire via transposable elements

Guillaume Bourque, Bernard Leong, Vinsensius Vega, Xi Chen, Yeng Ling Lee, Kandhadayar Srinivasan, Joon-Lin Chew, Yijun Ruan, Chia-Lin Wei, Huck-Hui Ng, Edison Liu

Genome Institute of Singapore, 60 Biopolis Street, #02-01, Genome, 138672, Singapore

Identification of lineage-specific innovations in genomic control elements is critical for understanding transcriptional regulatory networks and phenotypic heterogeneity. We analyzed, from an evolutionary perspective, the binding regions of seven mammalian transcription factors (TFs) identified on a genome-wide scale by different chromatin immunoprecipitation approaches and found that only a minority of sites appear to be conserved at the sequence level. Instead, we uncovered a pervasive association with genomic repeats by showing that a large fraction of the bona fide binding sites for five of the seven TFs (ER, p53, Oct4-Sox2 and CTCF) are embedded in distinctive families of transposable elements. Using the age of the repeats, we established that these repeat-associated binding sites (RABS) have been associated with significant regulatory expansions throughout the mammalian phylogeny. We validated the functional significance of these RABS by showing that they are over-represented in proximity of regulated genes and that the binding motifs within these repeats have undergone evolutionary selection. Our results demonstrate that transcriptional regulatory networks are highly dynamic in eukaryotic genomes and that transposable elements play an important role in expanding the repertoire of binding sites.

008: GEN2PHEN: an international effort to harmonise and optimise the databasing of gene-disease relationships

[1]**A. J. Brookes**, [2]D. Atlan, [3]C. Beroud, [4]E. Birney, [5]S. Brahmachari, [6]A. Cambon-Thomsen, [1]R. Dalgleish, [7]J. den Dunnen, [8]A. Devereau, [9]C. Diaz, [4]P. Flicek, [10]H. Gudbjartsson, [11]I. Gut, [12]T. Kanninen, [13]H. Lehvaslaiho, [14]J. Litton, [15]J. Muilu, [16]J. Oliveira, [4]H. Parkinson, [17]G. Patrinos, [18]G. Potamias, [19]E. Wingender, [20]L. Yip

[1]University of Leicester, Leicester, United Kingdom, [2]PhenoSystems S A, Lillois, Belgium, [3]Inserm, Montpellier, France, [4]European Bioinformatics Inst, EMBL, Hinxton, United Kingdom, [5]Council of Scientific and Industrial Research, Delhi, India, [6]Inserm, Toulouse, France, [7]Leiden University Medical Center, Leiden, Netherlands, [8]University of Manchester, Manchester, United Kingdom, [9]Fundació IMIM, Barcelona, Spain, [10]deCode Genetics, Reykjavik, Iceland, [11]Commissariat à l'Energie Atomique, Paris, France, [12]Biocomputing Platforms Ltd, Espoo, Finland, [13]University of Western Cape, Cape Town, South Africa, [14]Karolinska Institute, Stockholm, Sweden, [15]University of Helsinki, Helsinki, Finland, [16]University of Aveiro, Aveiro, Portugal, [17]Erasmus University Medical Center, Rotterdam, Netherlands, [18]Foundation for Research and Technology, Crete, Greece, [19]BioBase GmbH, Wolfenbuettel, Germany, [20]Swiss Institute of Bioinformatics, Geneva, Switzerland

With disease studies and genomics research producing ever more and ever larger datasets that connect genotypes and phenotypes, there is an urgent need for advanced informatics solutions that can handle this extensive and diverse information. Launched in January 2008, the GEN2PHEN project (Genotype-To-Phenotype Databases: A Holistic Solution) aims to help address this need.

The GEN2PHEN consortium (http://www.gen2phen.org/) involves 19 research and company partners; including 17 from Europe, one from India, and one from South Africa. Funding of 12 Million Euro from the European Commission (7th Framework Programme) is bolstered by additional funds provided by the partner institutions. Being a key European program, GEN2PHEN is intimately connected with other major related projects such as ENGAGE, CASIMIR, EUROGENTEST, BBMRI, ELEXIR, as well as the Human Variome Project (HVP).

The main objective of GEN2PHEN is to establish the technological building-blocks needed to evolve today's diverse databases into a seamless genotype-to-phenotype (G2P) biomedical knowledge environment, tied into genome browsers like Ensembl

The project's specific objectives include:

(1) Analysis of the current G2P informatics (2) Analysis of ethical aspects that need to be addressed (3) Development of key standards (4) Creation of generic database components and integration solutions (5) Creation of search modalities and data presentation solutions (6) Facilitation of data flows into G2P databases (7) Creation of a 'G2P Knowledge Centre' providing information exchange solutions, search/analysis tools, and support for primary data and comment deposition (8) Deployment of GEN2PHEN solutions to the community (9) Addressing questions of system durability and long-term financing

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 200754.

009: The HGNC database: an essential resource for the human genome

E. A. Bruford, S. M. Gordon, M. J. Lush, R. L. Seal, M. W. Wright

HUGO Gene Nomenclature Committee (HGNC), European Bioinformatics Institute, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton CB10 1SD, United Kingdom

The HUGO Gene Nomenclature Committee (HGNC) is an essential component of human genome management, and since 1989 has been the single authority for providing unique and user-friendly names and symbols for every gene in the human genome. Of the 25,000+ genes in our database most are protein-coding; we also name pseudogenes, phenotypic loci, some genomic features, and to date have named over 1,000 human non-coding RNA genes.

Approved gene symbols are based on names describing structure, function or homology wherever possible. Researchers are encouraged to contact the HGNC to request or confirm the approved nomenclature for specific genes and gene groupings, or to comment on the current gene names. Over 100 specialists advise us on the nomenclature of specific gene families, and we consult with our International Advisory Committee on policy issues.

Coordination with nomenclature committees for other species has proven invaluable; the HGNC has a very strong and active collaboration with the Mouse Genomic Nomenclature Committee (MGNC), which has proven essential in the parallel naming of orthologous human and mouse genes. In order to identify genes for orthologous naming we have developed our HGNC Comparison of Orthology Predictions search tool, HCOP (http://www.genenames.org/hcop) as a useful one-stop resource to summarise, compare and access various sources of orthology data. When sources agree on 1:1 orthologs between human and other mammals then these orthologs could directly adopt the human gene nomenclature. Indeed the Rat Genome and Nomenclature

Committee (RGNC) already aim to assign the approved human gene nomenclature for each orthologous rat gene.

The HGNC also has a strong working relationship with other databases including Entrez Gene, Ensembl, RefSeq, SwissProt/UniProt, Vega, UCSC, GeneCards and OMIM. All of these databases, and many more, prominently display HGNC gene symbols; using these symbols in an online search will then allow the user to retrieve information about the genes, including the structure and function of the encoded proteins, known genetic variation and clinical phenotypes, and related genes both in humans and in other species.

In 2007 we relocated to the European Bioinformatics Institute EMBL Outstation at Hinxton, near Cambridge in the UK, and our website can now be found at http://www.genenames.org/. For further information please visit the website or email us at mailto:hgnc@genenames.org. The work of the HGNC is supported by the NHGRI and the Wellcome Trust.

010: Flow of information in the *M. tuberculosis* interactome network: pathways to drug resistance

Nagasuma Chandra, Karthik Raman

Bioinformatics Centre, IISc, Indian Institute of Science, Bangalore, India

The global burden of tuberculosis has taken a new dimension in the recent years due to the emergence of drug resistant varieties of *Mycobacterium tuberculosis* MDR and XDR-TB, posing a major threat to TB eradication. Our ability to counter resistance is limited by a lack of understanding of how resistance emerges in bacteria. It is essential to understand the ways by which resistance can emerge upon exposure to a given drug. The reductionist approach of understanding proteins individually is obviously not sufficient, even at atomistic levels, making systems biology approaches essential to gain holistic insights. To a protein-protein interactome of *M. tuberculosis*, drug-induced expression data from literature were incorporated. The resulting network was analysed using computational approaches, to identify high propensity routes that would be traversed to bring about drug resistance. These routes form pathways from the drug targets to the proteins involved in extrinsic and intrinsic resistance mechanisms. Identification of these pathways forms the basis for a novel rational way to counter drug resistance. Our analysis shows that different targets are prone to resistance to different extents through different mechanisms. We introduce the concept of 'co-targets', which when simultaneously inhibited with the intended target, is likely to help in combating drug resistance. Different target-'co-target' pairs are identified in the study, which are expected to be useful in the design of new antitubercular drugs and to render existing drugs more useful. This approach is inherently generic, likely to significantly impact drug discovery.

011: Use of molecular modelling and QSAR for the design of beta-lactamase inhibitors

Dipan Chatterjee, Dipankar Chaudhuri

Department of Biotechnology, Heritage Inst of Technology, Chowbaga Road, Anandapur, Kol-700107, West Bengal, India

Beta-lactam antibiotics inhibit cell wall production in some bacteria by binding to DD peptidase, an essential enzyme for cell wall synthesis. This class of enzymes comprising beta-lactamase cleaves the beta lactam ring present in the beta-lactam class of antibiotics rendering them ineffective. The preferred medical solution has been to administer a concoction of a new class of beta lactamase inhibitors in combination with beta lactamase antibiotics. This research project is aimed at using molecular modelling and QSAR techniques specifically (and bioinformatics in general) to design new inhibitor molecules by first identifying the beta-lactam unit as the core structural unit of the lead compound. Lead optimization techniques were employed to introduce structural changes around this core and the efficacy of such modifications for utility as a drug were quantified through analysis of physicochemical parameters. These inhibitor molecules were designed based on non-covalent interactions such as electrostatic and hydrogen bonding necessary for tertiary structure stabilization of the complex between these molecules and the active site residues of beta lactamase. Modified molecules were docked with TEM-1 beta-lactamase using MOLEGRO virtual DOCKER. Parameters such as IC50, lipophilicity (log P), free energy of binding, steric effects and agreement with Lipinski's rule of 5 were used to rank these inhibitors for target binding and suitability as drugs. Docked complexes involving modified inhibitors had much stronger non-covalent interactions and much lower free energy minima satisfying the initial lead modification criteria of serving as effective inhibitors on the basis of enhanced binding affinities. As a representative example, the docked complex of modified 1-DiTrimethyl Amine clavulanic acid with beta-lactamase had a minimized free energy value of -101.2 kJ/mol and an IC50 of 188nM. The same modified inhibitor agreed with all of Lipinski's rule of 5 criteria for high success probability as a drug (e.g., Log P $= -$ 2.332; molar refractivity $=$ 80.2 etc.). A set of 10 such QSAR modified compounds were similarly ranked for binding mode efficacy, specificity and suitability as a drug. The ultimate goals of this project include developing a library of modified inhibitors using QSAR and molecular modeling principles and then subjecting them to vHTS (virtual High Throughput screening) software methodologies to rank these inhibitor compounds for suitability as a better class of beta lactam antibiotics.

012: Understanding Wnt cascades through modularization

Rajat De, Losiana Nayak

Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, India

Wnts are secreted cysteine-rich, lipid-modified glycoproteins. Wnt signaling plays crucial role in regulation of cell fate determination, proliferation, differentiation, migration, body plan formation, chondrocyte differentiation, limb initiation, bone growth, neural plate formation, and apoptosis. The pathway diversifies into β-catenin, planar cell polarity and Wnt/Ca^{2+} cascades, respectively.

Abnormal expression of Wnt family members result in various skin, bone and nervous disorders. For example, loss of LRP5 function causes OPPG syndrome while a gain-of-function mutation is reported for the HBM trait. Missense mutations in LRP5 gene lead to Van Buchem disease, autosomal dominant sclerosteosis and Osteoporosis type I syndrome. So analysis of Wnt cascades is necessary. But it is difficult to study these cascades as a whole due to its high degree of interconnectivity.

Hence we try to divide the human Wnt cascades using a modularization algorithm already developed by the authors, and study the different modules from biological perspective. Modularization is a process of splitting a network into sub-networks. Here human Wnt signaling pathway is subjected to be modularized into simple entities known as modules. A module can be defined as a subset of the original network that tends to be self-sufficient and has minimal dependency on the rest part of the network. The algorithm takes care of the fact that the created modules try to signify one or a distinct part of one function of the pathway. The input data is taken from KEGG/Pathway database.

The human Wnt signaling pathway gets modularized into 8 modules. The canonical Wnt pathway gets divided into modules WNT16, (DVL1)1, AXIN1, CTNNB1 and LEF1. Module WNT16 includes the Wnt ligand-receptor interactions in presence of co-receptors, several inhibitors and add-ons. Modules (DVL1)1, AXIN1 and CTNNB1 encounter for the steps of β-catenin balance and its entry into nucleus. Module LEF1 counters for further downstream actions that happen inside nucleus and effect expression levels of many concerned genes. Module (DVL1)2 is made up of the Planar cell polarity pathway and module PLCB1 is comprised of the Wnt/Ca^{2+} pathway. We cannot associate any biological significance to the very small Tp53 module. Further biological investigations of biomolecules present in these modules may throw light on cure of the diseases they cause. Modularized study of Wnt cascades of other species is also done to get an evo devo view.

013: Identification of phosphorylation sites and molecular modeling of DICER

Priyanka Dhar, Sayak Ganguli, Abhijit Datta

Bioinformatics Infrastructure Facility, Presidency College, 86/1, College street, Kolkata -700073, India

RNA interference (RNAi) is a post-transcriptional process initiated by double-stranded RNA molecules that induce degradation of a complementary target RNA. In the first step of the pathway, long double-stranded RNA molecules are chopped into shorter duplexes with 2 nucleotide overhangs at both 3′ ends by an endonuclease dubbed DICER. Protein phosphorylation is known to play a key role in regulating a variety of processes in eukaryotes, from the cell division cycle to neuronal plasticity. The most commonly observed phosphorylations affect serine, threonine and tyrosine residues although phosphorylation of histidines and aspartates has also been reported. Protein phosphorylation is catalyzed by enzymes called protein kinases, which are usually specific for either tyrosine or serine/threonine, with few of them being able to modify all three residues indistinguishably. The human genome encodes 518 protein kinases and recent estimates suggest that around one-third of cellular proteins could undergo phosphorylation. This work focused on the identification of phosphorylation sites in the amino acid sequences of DICER and proceeded towards comparative modeling of six important model organisms. The results obtained indicate that serine residues were the predominant phosphorylation target in all the samples. The search for conserved domains reveals that DICER has five different conserved domains along its entire sequence. Molecular modeling reveals that the structure of DICER resembles the paws of a crab with the groove forming the site of attachment of the target dsRNA. The domain of unknown function in the DICER enzyme showed homology with the chromatin remodeling protein domains. This suggests that DICER apart from being predominantly a ribonuclease may also perform chromatin remodeling at some stage. As the Ramachandran analysis revealed that the models had 96% residues in the allowed regions we can conclude that the models were satisfactory and could be used for future ligand binding studies.

014: In silico designing of inhibitor against Tin2 for aging

[1]**Anita Dubey**, [1]Shuchi Smita, [2]Ashish Patel, [3]Shailendra Kr Gupta, [2] M. K. Verma

[1]The Bioinformatica Solutions Lucknow, H.No.- 1/8, Vijay Khand Gomti Nager, Lucknow (U.P.), India, [2]National Institute of Technology Raipur, G.E. Road, Raipur,492010 (CG), India, [3]IITR, Lucknow, Post Box No. 80, Mahatma Gandhi Marg, Lucknow (U.P.), India

Telomeres are repeated TTAGGG DNA sequences that stop natural chromosome ends from behaving as random breaks, which might activate DNA-damage, also provide the structural basis for solving the Telomere shortening or end replication problem. In mammalian cells, telomere-binding proteins TRF1 and TRF2 play crucial roles in telomere biology. TRF2 protects chromosome ends and TRF1 regulates telomere length. They interact with several other telomere regulators including TIN2, PTOP, POT1, TPP1 and RAP1 to ensure proper maintenance of telomeres this complex is termed as Sheltrin or Telosome. TRF1's complex with TIN2, PTOP, and POT1 and regulates telomere length, whereas TRF2 mediates t-loop formation and end protection. Our prime target to inhibit the formation of sheltrin in case of aging and allow its formation in case of cancer. Crystal structure's of every protein involved in Sheltrin is already available except of TIN2 which was selected as target (Trf1-interacting protein, genbank ID Q9BSI4), 3 domains were predicted by prodom for Tin2 and the structure of relevant domain has been modeled by homology modeling, Tin2 is responsible for the binding of Trf1 and Trf2 in telomeres so in case of aging Tin2 is inhibited by a artificial peptide designed, which is docked with Tin2 by vakser lab's GRAMM-X tool for protein-protein docking, blocking Tin2's binding to Trf1 and Trf2. This will result in inhibition of sheltrin formation, allowing telomerase to continue replication of telomeres.

015: Unraveling the regulatory code of myelin genes

[1]**Debra Fulton**, [2]Eric Denarier, [3]Claire Tuason, [3]Hana Friedman, [1]Wyeth Wasserman, [3]Alan Peterson

[1]Centre for Molecular Medicine and Therapeutics, UBC, Vancouver, British Columbia, Canada, [2]Grenoble Institut des Neurosciences, Centre de Recherche Inserm U 836, Grenoble, France, [3]Molecular Oncology Group, McGill University, Montreal, Quebec, Canada

The myelin sheath is an insulating layer that wraps around the thread-like axonal extensions of neurons to enable rapid, efficient transmission of impulses along nerve fibers. Two types of glial cells are responsible for the biogenesis and elaboration of myelin from their cell body plasma membranes: Schwann cells in the peripheral nervous system (PNS) and oligodendrocytes in the central nervous system (CNS). The proteins produced during the assembly of myelin, along with their encoding genes (herein referred to as myelin genes), are responsible for its unique structural and functional properties. Myelin-related gene expression appears to be largely controlled by transcription factors (TF); only a portion of which are known. Dysregulation of myelin sheath production or maintenance is associated with multiple human diseases, such as multiple sclerosis and schizophrenia.

We identified prospective regulatory regions using sequence conservation of non-coding DNA neighboring orthologous myelin-related genes. Each region was prioritized based on the prediction of co-located motif instances that could potentially interact with a myelin-regulating TF. Using reporter constructs and a controlled strategy of transgenesis in mice, the prioritized sequence regions were tested for in vivo function in oligodendrocytes and/or Schwann cells. The sequence segments that were demonstrated to be functional, along with their temporal parameters, are being used to seed an in silico-based analysis for the identification of sequence features and characteristics which could be responsible for the spatio-temporal transcriptional events of co-expressed sets of myelin genes. This in silico analysis incorporates gene expression data mining and analyses to expand the myelin gene set and establish spatio-temporal groupings, along with the identification of statistically relevant regulatory sequence feature signals emanating from the co-expressed sets of myelin genes.

016: A novel transmission disequilibrium test for quantitative traits

Saurabh Ghosh

Indian Statistical Institute, Human Genetics Unit, 203 B T Road, Kolkata 700108, India

The classical transmission disequilibrium test (TDT) for binary traits proposed by Spielman et al. (1993) is a family-based alternative to population-based case-control studies and circumvents the problem of population stratification as it tests for allelic association in the presence of linkage. However, since the clinical end-point traits are often defined by quantitative precursors, it has been argued that it may be a more prudent strategy to analyze the quantitative phenotypes without dichotomizing them into binary traits. The paradigm of linkage disequilibrium in the context of quantitative traits generally considers the intuitive concept of differences in allelic frequencies between individuals having high values of the quantitative trait and those with low values of the trait as evidence of linkage disequilibrium between the marker locus and the QTL. While Analysis of Variance (ANOVA) is a popular approach for association analyses of population-based quantitative trait data, it suffers from the inherent problem of population stratification, and hence, it is of interest to explore for family-based association methodologies using transmission patterns. Although some methods have been developed for testing transmission disequilibrium in the context of quantitative traits, these are not direct extensions of the classical TDT. We propose a simple logistic regression based test that can be analytically shown to be statistically equivalent to the TDT for binary traits, and hence is not susceptible to the presence of population stratification in the data. We perform Monte-Carlo simulations under a wide spectrum of disease models and varying parameter values of linkage disequilibrium to evaluate the power of the proposed procedure. We find that similar to the binary TDT, the power decreases with increase of dominance and decrease of heterozygosity at the QTL. The proposed method can be easily extended to incorporate multivariate phenotypes. We apply our method to analyze externalizing symptoms, an alcoholism related endophenotype from the Collaborative Study on the Genetics of Alcohism (COGA) project.

017: Dynamic changes in protein functional linkage networks revealed by integration with gene expression data

S. R. Hegde, P. Manimaran, S. C. Mande

Centre for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad-500 076, India

The pattern of interactions among various biomolecules exhibits dynamic behavior, thereby enabling cells to react to many changing conditions. Proteins being the dominant macromolecules that carry out routine cellular functions, understanding the dynamics of protein:protein interactions might yield useful insights into the cellular responses. The large-scale protein interaction data sets are, however, unable to capture the changes in the profile of protein interactions. In order to understand how these interactions change dynamically, we have constructed conditional protein linkages for *Escherichia coli* by integrating genome-wide functional linkages and gene expression information. The gene expression information was retrieved from microarray data available publicly. As a case study, we have analyzed UV exposure in wild-type and SOS deficient *E. coli* at 20 min post irradiation. The sub-networks thus constructed are hypothesized to represent a real functional interaction picture of the cell. We have further applied various graph theoretical measures to extract the relevant biological information from these sub-networks. These conditional networks exhibit overall robustness against external perturbations as revealed by the analyses of their topological properties. Although the global topological properties of the networks are similar, many subtle local changes are observed, which are suggestive of the cellular response to the perturbations. Some such changes correspond to differences in the path lengths among the nodes of carbohydrate metabolism correlating with its loss in efficiency in the UV treated cells. Similarly, expression of hubs under unique conditions reflects the importance of these proteins. One such hub, the DNA replication inhibitor Hda, is found to be expressed only upon UV treatment. We observe the unique expression of the genes involved in the iron uptake system in untreated wild-type cells. Various centrality measures applied on the networks indicate increased importance for replication, repair and other stress proteins for the cells under UV treatment, as anticipated. We are able to identify many repair proteins such as DinG, DnaN, MutM, MutS, RuvC and RecF critical for the UV treated networks in terms of degree centrality. This we propose to be a novel methodology in which a raw microarray data can be analyzed by incorporating molecular interaction information with gene expression.

018: 'Sexual dysfunction' in the brain: an uncharacterized gene's apparent role in male erectile dysfunction

[1]**Mustak Ibn Ayub**, [2] M. D. Maksudul Alam, [1]S. M. Mahbubur Rashid, [1]Mahdi Muhammad Moosa

[1]Department of Genetic Engineering and Biotechnology, University of Dhaka, Dhaka-1000, Bangladesh, [2]Department of Biochemistry and Molecular Biology, University of Dhaka, Dhaka-1000, Bangladesh

Erectile dysfunction in man has been correlated with physical and mental conditions. Though the physical causes are well defined, the mental roles are yet a mystery. Here, we propose the function of an uncharacterized gene, UBAP2 which expresses predominantly in the brain and apparently has a key role in male erectile dysfunction. Bioinformatics analyses show that the UBAP2 protein plays a pivotal role to antagonize the function of Rho/Rho kinase and thus protect eNOS (endothelial Nitric Oxide Synthase) dependent penile erection pathway. It has already been proven that the eNOS mediated signaling pathway determines the erection strength and time during sexual activities. So, if UBAP2 gene is mutated (or abnormal), the activity of Rho/Rho kinase increases and eNOS fails to carry out its normal function which results in erectile dysfunction. Relevant bioinformatics analyses have confirmed this prediction.

019: Genetic variants of constitutive androstane receptor (CAR) gene and the diseases in the Japanese elderly population

[1]**Shinobu Ikeda**, [2]Tomio Arai, [3]Makiko Mieno, [1]Masaaki Muramatsu, [2]Motoji Sawabe, [4]Noriko Tanaka

[1]Department of Molecular Epidemiology, Medical Research Institute, Tokyo Medical and Dental University, 2-3-10 Kanda-Surugadai, Chiyoda-ku, Tokyo, Japan, [2]Department of Pathology, Tokyo Metropolitan Geriatric Hospital, 5-2 Sakaecho, Itabashi-ku, Tokyo, Japan, [3]Department of Medical Informatics, Center of Information, Jichi Medical University, 3311-1 Yakushiji, Shimotsuke-shi, Tochigi-ken, Japan, [4]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, United States of America

Background: Constitutive androstan receptor (CAR) encoded by NR1I3 is a member of the orphan nuclear receptor family. CAR regulates a number of genes encoding enzymes involved in xenobiotic/endobiotic metabolism, conjugation, and transport of small hydrophobic substrates. Previous studies have indicated that CAR participates in bilirubin clearance, bile acid detoxification, and adaptive responses to nutrition stress, demonstrating the physiological importance of CAR. We previously identified 4 novel rare variants in Japanese population, and performed functional analysis of these variants. Two rare variants caused marked reduction in transactivation of the reporter gene and in the response to a human CAR-specific agonist. Thus, variations of CAR resulting in functional alterations may have serious impacts on pharmacological and physiological responses of the target genes such as UGT1A1 and CYP3A4. However, there is no information on the association between CAR polymorphism and human diseases.
Methods: In this study, we selected 9 polymorphisms including rare variants in NR1I3 gene and genotyped them in consecutive autopsies of elderly Japanese population registered in the JG-SNP (The Japanese SNPs for geriatric research) database using LightCycler480 system (Roche Diagnostics). Multivariate logistic regression analysis was done to calculate association between each polymorphism and geriatric 26 diseases or 12 pathological-identified diseases and conditions. Each polymorphism was assessed according to dominant, recessive, and additive genetic models and adjusted for risk factors. A p-value lower than 0.05 was considered statistically significant. All statistical analyses were performed with SAS statistical software ver.9.1.3 (SAS Institute Inc., NC).
Results: The rare variants were not detected in this population. However, we identified a novel rare variant +760C > T, a nonsense mutation which leads to Agr254Stop. The three carriers were patients of liver cancer, prostate cancer, and lymphoma. A common variant, +540C > T was associated with hematopoietic malignancy and leukemia. After adjusting for potential confounders (age, gender smoking and drinking) the risks were significantly lower in carriers of two variant alleles (OR; 0.59, 95% CI; 0.38–0.91, and OR; 0.48, 95% CI; 0.23–0.98, respectively).
Conclusion: Polymorphism of CAR, which controls the metabolism and elimination of endogenous and exogenous toxic compounds, may play an important role in the development of hematopoietic malignancy.

020: Target identification by in-silico studies on NS1 proteins of bird flu virus

Nandan Kumar Jana, Amit Singh, Arkadeep Sinha, Arghya Sett, Gourab Chatterjee

Department of Biotechnology, Heritage Institute of Technology, Chowbaga Road, Kolkata-700107, WB, India

Avian influenza, or 'bird flu', is a contagious disease of animals caused by Influenza viruses that normally infect only birds and, less commonly, pigs. Avian influenza viruses are highly species-specific, but have, on rare occasions, crossed the species barrier to infect humans. There are several different types of avian viruses (type A, B, C with each type having several subtypes) and they are constantly evolving. And this ability of influenza viruses to change their genetic makeup and to swap genes makes them unpredictable and potentially deadly. In this work the focus was on the translated NS1 protein of the H5N1 virus (as it is responsible for pathogenesis and elevation in levels of cytokines during infection) along with a comparative study of H5N1's translated NS1 protein with the translated NS1 protein of the other pandemic virus (H1N1, H2N2, H3N2) infecting humans, avian and swine. Attempt has been done to determine and create the 3D structure of the consensus protein sequence translated from the Nonstructural NS1 gene within the host cells by H5N1 flu viruses. After comparing the NS1 protein sequence of H5N1 along with the other three pandemic viruses' translated NS1 protein sequence, a common region/domain has been identified which seems to have remain conserved in all pandemic flu viruses including H5N1 subtype. A drug/vaccine can be devised to attack this conserved site as this site is present in all pandemic flu viruses which is unaffected by mutations in the viruses. So the drug/vaccine can be more effective in combating the H5N1 virus if it interacts with this conserved region thereby decreasing the severity of the disease and mortality rate.

021: Meta-analysis of condition-specificity of gene expression: an integrated statistical framework for analysing 20,000 + transcriptomics assays across multiple studies and platforms in ArrayExpress

[1]**Misha Kapushesky**, [1]Helen Parkinson, [1]Wolfgang Huber, [2]Michael Ashburner, [1]Alvis Brazma

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, United Kingdom, [2]Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, United Kingdom

Gene regulation pathways have been shown to be active simultaneously in numerous systems in living cells, functioning in different organs and under a diversity of conditions [1]. It is hypothesized, and there exists strong evidence in individual cases (e.g., [2]), that the expression of some regulatory elements and targets of pathways is restricted to specific tissues or conditions or is specifically required in certain tissues in developmental processes. We present in this study an analytical framework where the wealth of high-throughput gene expression data can be examined cohesively in order to test this hypothesis and related questions.

Using a robust statistical framework (based on the limma [3] package in Bioconductor), we tested the differential expression strength of genes across more than 2,000 conditions studied in over 20,000 assays in 9 species (including human, mouse, rat, fruit fly and two yeasts). Furthermore, we developed a means to measure the confidence of specificity of expression of a given gene in a condition, including tissue, disease, cell type, and developmental stage. We present here an analysis of the distribution of condition-specific active genes, elucidating genes whose expression patterns tend to be determined by few conditions as opposed to those expressed without significant condition association. A method and results of its applications are presented for using the obtained global statistics for studying functionally related groups of genes (e.g., pathways) to discover core subsets whose patterns of condition-specific expression are similar.

The resulting resource is available online (http://www.ebi.ac.uk/microarray-as/atlas) as the ArrayExpress Atlas of Gene Expression, a database of summary statistics over a curated set of public microarray data from the ArrayExpress and GEO repositories [4, 5]. A simple interface to the complex statistics allows biologists and non-expert bioinformaticians to get a meaningful overview where and under which conditions a gene (or a group of genes) of interest is active.

References

[1] Matushansky I (2008) Cell Cycle, 2008 Mar; 7(6):720–4
[2] Ober EA et al (2006) Nature, Aug 10; 442(7103):688–91
[3] Smyth GK (2005) In: Gentleman R et al (eds) Bioinformatics and computational biology solutions using R and bioconductor, Springer, New York

[4] Parkinson H et al (2007) Nucleic Acids Res, Jan; 35:D747–50
[5] Barrett T et al (2007) Nucleic Acids Res, Jan; 35:D760–5

022: Assembly-based analysis protocol of metagenomic short-read sequence data

Shinji Kondo, Todd Taylor

MetaSystems Research Team, Computational Systems Biology Research Group, RIKEN Advanced Science Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan

Reduction of sequencing costs is imperative for metagenomic research, which requires a huge amount of sequence data delivered from uncultivatable mixtures of a number of microbial species. Although sequencing by synthesis (Illumina/Solexa and SOLiD) can generate a gigabase or more of data (1/3 of the human genome) in a single operation, their short read lengths (20–40 bp) are not suited to conventional sequence analyses. To overcome this shortcoming, we wished to develop an analysis method based on assembly of the short reads. Using simulated short-read (36 bp) sequence sets which include genomes from a number of multi-strain bacterial species and also varying amounts of 'noise' sequences, we have tested the capability of a published (single-genome) assembly program (Velvet, Zerbino and Birney 2008, Genome Research) to assemble contigs large enough to allow subsequent sequence analyses. When a certain sequence depth was achieved in the synthetic metagenomic sequence data, Velvet generated a number of accurate contigs large enough to be used for metagenomic analyses. We measured dependence of the contig number and size distribution on the sequence error rate, sequence depth and insert size for paired-end reads and then computed the threshold values required for phylogenetic binning and gene prediction using the assembled contigs. Although our approach has yet to be tested with real short-read sequence data, our simulations suggest that this cost-effective assembly-based method is feasible, particularly for metagenomic samples such as human gut which contain a relatively small number of (dominant) species resulting in a high level of sequence redundancy.

Reference

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829

023: Enhancements in widely used primer design program Primer3

Triinu Koressaar, Maido Remm

University of Tartu, Riia 23-305, Tartu 51010, Estonia

Primer3 is widely used open source software for primer design. The program was originally written more than a decade ago. Despite that, this program is maintained and developed constantly and thereof is becoming one of the central freeware primer design programs. We have introduced several enhancements into the Primer3 (version 1.1 and newer). These comprise new formula for calculating melting temperature with two new different formulas for monovalent salt correction which together improve melting temperature predictions approximately 10°C compared to Primer3 not having these enhancements. Moreover, the average differences between the experimental melting temperature and melting temperature predicted by updated Primer3 is 1.37°C. Also the updated Primer3 can take into account effects of divalent cations which are included in most PCR buffers. Formation of undesirable monomer and dimer structures of sequences of primers and sequences of products may lead to PCR failure. Additionally, predictions of occurrences of primer-dimer and -product interactions, primer- and product-monomer structures based on thermodynamical stability are added into the Primer3.

024: GeneScan: a context independent gene finding program

Kousik Kundu, Shrish Tiwari

Centre for Cellular and Molecular Biology, Uppal Road, Habsiguda, Hyderabad 500 007, AP, India

Computational gene finding is still an open problem of high relevance. Many algorithms have been developed over the years, but most of them depend on the availability of a set of genes for training their program. Here we present an algorithm, GeneScan, which is context independent, robust to sequencing errors and fairly simple to implement. It is based on the observation of short-range correlations, specifically a 3-periodicity, in protein-coding genes, which is absent in non-coding DNA. GeneScan uses the mathematical technique of Fourier transform to compute correlations between nucleotides of a DNA sequence. The basic aim of this study is to develop an algorithm which is highly specific, i.e. with very few false positives, that can be applied to the analysis of any genome, without prior training, and which can be used on draft sequences, without significantly affecting its prediction accuracy. The algorithm has presently been tested extensively on prokaryotic genomes and we are extending it to work for eukaryotic genomes as well. The performance of GeneScan is comparable to existing methods. Sensitivity and specificity for the Escherichia coli for GeneScan are 87% (GLIMMER—97%) and 91% (GLIMMER—90%), respectively.

025: Softwares and databases developed for innate immunity

Sneh Lata, G. P. S. Raghava

Institute of Microbial Technology, Sector-39A, Chandigarh-160036, India

In the decades gone by, research in the field of subunit vaccine designing has been directed mainly to fish out new subunit vaccine candidate from pathogens. Identification of potential vaccine candidates continued to be a major challenge for experimentalists, in subunit vaccine designing. For this reason a large number of bioinformatics databases and tools have been developed to predict the antigenic epitopes or antigens. However, only limited success has been achieved so far as the epitopes/antigens show poor immunogenicity (when tested experimentally), thereby defeating the main goal of vaccination i.e. to provide strong and long lasting immunity. Accumulating evidences now suggest that addition of components of innate immunity along with these epitopes helps to generate a strong and effective immune response and play a fundamental role in influencing immunological memory. Innate immune system employs a wide variety of receptors called Pattern-Recognition Receptors that detect evolutionarily conserved molecular patterns from pathogens called Pathogen Associated Molecular Patterns (PAMPs). Stimulation of Pattern Recognition Receptors by their ligands leads to activation of adaptive antigen-recognition receptors and induction of key co-stimulatory molecules, antimicrobial peptides and cytokines as well as maturation and migration of other cells. Overall, these cascade of events leads to the development of a robust and durable adaptive immune response. Thus, the PAMPs can be used as adjuvants to be administered along with

subunit vaccines in order to enhance its efficacy. Understanding the molecular mechanisms responsible for recognition of PAMPs and generation of downstream signaling effecter molecules therefore, would be crucial for the development of new approaches to vaccine formulation and immunotherapy. Although in the last decade progress has been made by the scientists to understand the innate immune system but bioinformatics based research for innate immunity is still in its infancy. We have developed at IMTECH a database and few servers to predict the crucial molecules of innate immune system. The database PRRDB (http://www.imtech.res.in/raghava/prrdb) houses the information about PRRs and the PAMPs; the server antibp (http://www.imtech.res.in/raghava/antibp) predicts the antibacterial peptides whereas the web server cytopred (http://www.imtech.res.in/raghava/cytoprd) predicts the cytokines as well as their families and sub-families.

026: Obesity, diet and type II diabetes risk in Indians

[1]**Terresa Lehman**, [1,2]Luke Ratnasinghe, [1]Rama Modali, [1]Mike Seddon, [3]K. Nasaruddin, [3]J. B. Vijayakumar, [4]Charles J. Spurgeon, [4]K. Radha Mani, [4]Seema Bhaskar, [4]P. Smitha, [4]Giriraj Chandak

[1]BioServe Biotechnologies, Ltd., Beltsville, MD, 20705, United States of America, [2]Genomic Nanosystems LLC, Beltsville, MD-20705, United States of America, [3]Bioserve Biotechnologies (India) Ltd, Hyderabad, India, [4]Center for Cellular and Molecular Biology, Hyderabad, India

Incidence of type II diabetes is increasing in India. Obesity, diet and exercise are modifiable risk factors for type II diabetes. We evaluated the association between obesity and risk of diabetes among 3,369 type II diabetes cases and 2,687 non-diabetes controls (total n = 6,056) in the BioServe global epidemiology study (GES). We also evaluated the association between recently discovered single nucleotide polymorphisms and risk of diabetes in a nested case-control study of 360 cases and 360 controls. The GES is a multi-national study to assess disease risk factors and is linked to the Global Repository that houses biomaterial including DNA. For diabetes, newly diagnosed subjects provided informed consent and were asked about health behaviors using a validated survey instrument. Indian diabetes cases consumed less vegetable per day than controls ($p < 0.001$). Furthermore, Indian diabetes cases consumed more fish and red meat than controls ($p < 0.02$). The mean body mass index (BMI) among diabetes cases was higher than that among controls ($p < 0.01$). We used multivariate unconditional logistic regression to estimate the association between BMI and risk of diabetes among South Indians, the multivariate Odds Ratio (OR) was 2.10 (95% Confidence Interval (95% CI): 1.68–2.63) for individuals who were obese (BMI > 30 kg/m^2) compared to healthy-weight (BMI 18.5–24.9 kg/m^2) individuals after adjusting for age, gender and pack-years of smoking. Among North Indians the association between obesity and diabetes risk was similar to that among South Indians (OR: 2.00, 95% CI: 1.02–3.90). We will also report the association between 20 SNP genotype markers and risk of type II diabetes. Results from our study strongly confirm that maintenance of healthy weight and diet are important public health messages in the effort to reduce increasing incidence of type II diabetes.

027: Expression divergence during human evolution is shaped by change in genomic neighbourhood of genes

M. Madan Babu, Sarah Teichmann, Subhajyoti De

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB20QH, United Kingdom

Mutations that alter the expression level or expression pattern of genes can contribute to the evolution of new species. Such changes may arise due to small-scale mutations that affect a single or a few base pairs, or due to large-scale events such as segmental duplication or chromosomal re-arrangements. Although the contribution of such mutations to the evolution of gene expression pattern has been well studied, the role of intermediate scale mutations that alter the genomic neighbourhood of a gene, i.e., position effect, remains largely unexplored. Such mutations which affect the neighbourhood of a gene may be a result of (1) recombination or duplication event that resulted in the incorporation of the gene into a completely different region or (2) insertion or deletion of genetic material around the gene.

In this work, we investigate if a change in the genomic neighbourhood of orthologous genes between human and chimp has resulted in expression divergence. By comparing genomic neighborhood of orthologous genes from human and chimp with the expression levels of transcripts from several equivalent tissues, we demonstrate that genes with altered neighborhood are likely to undergo significant expression divergence. In addition, a detailed analysis revealed that the genes which are highly expressed in the prefrontal cortex and several other brain tissues show a high incidence of altered gene neighborhood when compared to its chimpanzee ortholog. Although some of these genes were precisely characterized in the context of disease associated with human cognition, an excess of cognition related genes with altered neighborhood suggests that such a mechanism might have played an important role in the evolutionary path leading from the primate ancestor to human. Taken together, our findings suggest that, in addition to other molecular processes, position effect induced gene expression change is a fundamental mechanism that contributes to transcriptome evolution and ultimately to the evolution of new species.

028: Predicting signaling networks upon inflammation caused by PAR-1 activation: application of the ExPlain™ analysis platform

[1]**Dinesh Madhyastha**, [1]Neetu Tandon, [1]Mangala Sathyamurthy, [2]Tatiana Konovalova, [2]Volker Matys, [2]Olga Kel-Margoulis, [2]Alexander Kel

[1]Biobase Databases India Private Limited, 32/1 Crescent Road, Bangalore-560 001, India, [2]BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany

Background: Activation of G-protein coupled receptor, PAR-1 (Protease Activated Receptor (1) by thrombin results in the rapid transcription of genes involved in inflammation. The aim of our study is to analyze molecular events leading to the differential expression of genes in mouse urinary bladder cells in response to inflammation caused by PAR-1 activation. We have applied the previously developed ExPlain™ Analysis Platform (BMC Bioinform, 7(Suppl. 2):S13, 2006) to microarray results published in (BMC Physiol, 7:3, 2007). Computational analysis was done via comparison of 75 differentially expressed genes (**Yes set**) versus genes with a different tissue-specificity (**No set**).

Results and Conclusions: The bioinformatics analysis performed included several steps. First, 75 differentially expressed genes were classified by MESH terms according to categories therapeutic target, biomarker, and molecular mechanism (HumanPSD™ Disease View, Nucleic Acids Res, 2002, 30:137). Several of the upregulated genes were found to be associated with ovarian, pancreatic and colorectal neoplasms, including AKT2, DUSP1, ELK1, HMGA1, PTPN1, JUND, NFKBIA, among others. Several genes were also found to be involved in HIV infections. Next, we mapped putative transcription factor (TF) binding sites in the promoters of our differentially

expressed genes to find hits overrepresented in PAR-1 dependent genes. This analysis was based on the TRANSFAC® library of positional weight matrices (Nucleic Acids Res, 34:D108, 2006). At the next step, promoter models were constructed for the differentially expressed genes using the Composite Module Analyst program in ExPlain. TFs suggested to be involved in coordinated regulation of differential expressed genes include CEBPD, EGR1, EGR2, EGR2, and NFKB1. Finally, we analyzed networks upstream of our suggested TFs to reveal key nodes that might be responsible for coordinated regulation of the set of TFs. Our key node search algorithm is based on the signaling and metabolic networks collected in TRANSPATH® (Nucleic Acids Res, 34:D546, 2006). Key node analysis allows us to hypothesize PIP3, Vav1, Zap-70 as potential targets which are activated upon PAR-1 stimulation. Analysis workflow, functional classification of the differentially expressed genes, detailed promoter models with putative TF binding sites as well as most prominent key nodes and signaling networks will be presented on the poster.

029: Omega parameters: a novel integrated approach for feature detection in nucleic acid structures

[1]**Abhijit Mitra**, [1]Swati Jain, [1]Raina Arora, [1]Nagarjuna Kr, [2]Dhananjay Bhattacharyya

[1]Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology Hyderabad, Gachibowli Hyderabad 500032, India, [2]Biophysics Division, Saha Institute of Nuclear Physics, 1/AF Bidhannagar Kolkata 700064, India

DNA structures and their variations are important in several contexts such as chromatin remodeling and transcriptional initiation and regulation etc. The need for computational characterization of DNA structures associated with DNA bending/binding, B - > A transitions in protein-DNA complexes, molecular dynamics trajectory analysis and also in the context of strand orientation, especially in DNA triplexes and quadruplexes, have given rise to several fine grained parameters, such as backbone conformations, sugar pucker and χ angle, local strand orientation, base pair and base pair step parameters, backbone base inclination etc. With their unpaired bases, extensive non-helical regions, non-canonical base pairs and tertiary interactions, RNA molecules host a large variety of recurrent structural and functional motifs and parameters such as zp, defined in the context of double helical DNA base pairs and base pair steps and effectively used for A and B DNA differentiations, can not be used for computationally driven automated mining of complex RNA motifs. Thus for RNA structures, though several motif mining algorithms based on backbone conformational parameters and involving computational detection and analysis of base pairs and their parameters are available, there are no reports of studies involving parameters which simultaneously involve backbone conformation as well as base pairing geometry. In this work, we propose a novel set of parameters, Omega Torsions and Omega Distances, which captures backbone geometries at base(i)—base(j) interaction points and which can be used to computationally characterize structural features both in RNA and DNA. Pseudo torsion angles: Omega η: P(i)–C4′(i)-C4′(j)–P(j); Omega θ: P(i + 1)–C4′(i)–C4′(j)–P(j + 1) Omega 1: P(i)–C4′(i)-C4′(j)–P(j + 1); Omega 2: P(i + 1)–C4′(i)–C4′(j)–P(j) Pseudo bond distance: Omega distance: C4′(i)–C4′(j) Where 'i' and 'j' are bases interacting through hydrogen bonds. We report the results of our benchmarking studies of Omega parameters, as applied to protein DNA complexes including e.g., TATA box-TBP complexes and nucleosomal DNA, with standard DNA structure classification parameters, such as zp, and of our preliminary studies on their effectiveness in classifying local geometries of RNA strands in different base pairing contexts. We conclude that omega parameters can potentially lead to the development of a single set of parameters which can be used to characterize both DNA as well as RNA structures.

030: GC-rich special short sequences dominate eukaryotic promoter regions

[1]**Chanchal K. Mitra**, [2]Luciano Milanesi

[1]University of Hyderabad, P.O. Central University, Hyderabad 500 046, India, [2]CNR-ITB Institute of Biomedical Technologies, Segrate (Milan), Italy

We report the presence of a small group of penta- and hexanucleotides around the promoter region of the human genome. We have downloaded the eukaryotic promoter sequences from http://www.epd.isb-sib.ch/seq_download.html with redundancy removed. We have subjected the 1796 sequences to pairwise alignment, without gaps, with the restriction that the shortest aligned sequence must have a length of five or more. The common sequences were sorted and unique sequences were counted. The sequences were next ordered according to their frequency. This was plotted as a histogram and we immediately note that the graph is far from uniform.

If we assume that the four NT bases are all equally likely, we shall be getting all possible 5-nt sequences with a frequency of once every 1024 and 6-nt sequences once every 4096 bases. This has not been observed. We have also noted the actual position of occurence of the more common sequences and see that the common subsequences do not occur exactly at the same place. We also note that a number of subsequences are seen on the coding side of the TSS. This is unexpected but may have some role in the early processes of transcription. It may not be fair to assume an equal population of the 4 NTs around the TSS. We have therefore computed the proportions of A + G and A + T along the sequence and see some interesting pattern. The features are broad but clearly unmistakable. The sharp peak at the TSS is caused by the start codon. We conclude that there exists a relatively few number of conserved sequences at/around the TSS and this includes both downstream and upstream regions. We also conclude that these sequences are clustered around some preferred regions. We have listed a few of the most common sequences (5- and 6-NTs). These need to be studied in more detail.

031: Analysis of regulatory motif binding site: an effective approach for identifying upregulated genes in the human oncogenome

[1]**Sanga Mitra**, [1]Arpita Basu, [1]Debipriya Banerjee, [1]Manjita Mazumdar, [1]Sambit Basu, [1]Sananda Chakraborty, [1]Satabdi Kundu, [1]Jyotirmoy Ghosh, [1]Subhankar Paul, [1]Satarupa Chatterjee, [1]Urmimala Sen, [2]Sayak Ganguli

[1]Indian Institute of Social Welfare and Business Management, Kolkata, 700073, India, [2]BIF, Presidency College, Kolkata, 700073, India

Current trends in the analysis of cancer genomics have involved approaches for identifying the expression levels of the oncogenome in specific types of cancer and to locate the genes that are upregulated in specific cancer phenotypes. The experimental protocols mostly involve SAGE and microarray based approaches, which are both time consuming and expensive. Moreover the analysis involves a huge amount of starting gene sets which may or may not respond evenly in each experimental cascade. In this work we report for the first time an in silico approach for the detection and screening of genes which might be upregulated in cancer. The approach involves the

identification and analyses of regulatory motif binding sites in the genes implicated in various human carcinomas and correlating them with the reported expression patterns of the genes. The results indicate that the genes reported to be upregulated in cancer possess a specific pattern of regulatory motif binding sites and future cancer biomarkers can be screened and tested with the presence of the same set of regulatory motif binding sites. A correspondence analysis was also performed along with a correlation analysis to further establish our findings. This approach could prove to be beneficial for screening of the specific genes before workers embark upon wet lab experiments.

032: Association tests using kernel-based measures of multilocus genotype similarity between individuals

[1]**Indranil Mukhopadhyay**, [2]Eleanor Feingold, [2]Daniel E. Weeks, [3]Anbupalam Thalamuthu

[1]Department of Statistics, University of Burdwan, Golapbag, Burdwan, West Bengal, 713104, India, [2]Department of Human Genetics and Biostatistics, University of Pittsburgh, 130 DeSoto Street, GSPH, Pittsburgh, PA 15261, United States of America, [3]Human Genetics/Computational and Mathematical Biology, Genome Institute of Singapore, 60 Biopolis Street, #02-01, Singapore 138672, Singapore

In a genetic association study, it is often desirable to perform an overall test of whether any or all single nucleotide polymorphisms (SNPs) in a gene are associated with a phenotype. Here we propose a new kernel based association test (KBAT) of joint association of several SNPs. Our test is non-parametric and robust. It can be used to test multiple correlated SNPs within a gene and can also be used to test independent SNPs or genes in a biological pathway. Our test uses an analysis of variance (ANOVA) paradigm to compare variation between cases and controls to the variation within the groups. The variation is measured using kernel function for each marker and then a composite statistic is constructed to combine the markers into a single test. We present simulation results comparing our statistic to the U-statistic based method by Schaid et al. and another statistic by Wessel and Schork. We consider a variety of different disease models and assumptions about how many SNPs within the gene are actually associated with disease. Our results indicate that our statistic has higher power than other statistics under most realistic situations.

033: Classifiers for genome annotation built on gene ontology

Soumyadeep Nandi, Andrew M. Lynn

Jawaharlal Nehru University, School of Information Technology, New Delhi, India

Sequence annotation is a classification problem. A large variety of methods (viz. sequence similarity based methods as BLAST; profile based method PSIBLAST, HMMER, etc.) are used to assign function to a sequence by comparing it with well annotated databases using the top hit annotation as the prediction for a query instance. Methods to classify sequences without known orthologs are still in development. Conserved hypothetical proteins—i.e. predicted proteins conserved in more than one organism constitute a substantial fraction of novel sequenced genomes.

Hierarchical classification systems and disparate sources of data can improve classification using supervised learning by providing more training information. Hierarchical classification systems provide two advantages: Firstly, the ability to build more discriminative classifiers using positive and negative training sequences and secondly to assign sequences at a lower (finer) level of hierarchy or precise functional category. Exploiting the hierarchical structure of Gene Ontology and annotating through GO terms provides us a better and precise annotation with the controlled vocabularies of Gene Ontology.

Supervised learning methods have been used earlier to classify sequences, though restricted to sub-families that share a significant homology. In our approach we use a support vector machine. The training data included sequences mapped onto the Gene Ontology terms, and the system trained using patterns of fold extracted from the pfam and superfamily databases as well as functional motifs using the PROSITE database.

We validate this system on a subset of the tree, whose leaf nodes are completely populated. The sensitivity and specificity were 92.2 and 95.1%, respectively. This system is used to functionally characterize proteins previously classified as ‘conserved hypothetical’.

034: Prediction of protein-protein interactions between a malarial parasite and human

Srinivasan Narayanaswamy, Krishnadev Oruganty

Molecular Biophysics Unit, Indian Institute of Science, India

Lack of large-scale efforts aimed at recognizing interactions between host and pathogens limits our understanding of many diseases. We present a simple and generally applicable bioinformatics approach for the analysis of possible interactions between the proteins of a parasite, Plasmodium falciparum, and human host. In the first step, the physically compatible interactions between the parasite and human proteins are recognized using homology detection. This dataset of putative in vitro interactions is combined with large-scale datasets of expression and sub-cellular localization. This integrated approach reduces drastically the number of false positives and hence can be used for generating testable hypotheses. We could recognize known interactions previously suggested in the literature. We also propose new predictions which involve interactions of some of the parasite proteins of yet unknown function. The method described is generally applicable to any host-pathogen pair and can thus be of general value to studies of host-pathogen protein-protein interactions.

035: Genomic control for fisher’s exact test

Yukinori Okada, Ryo Yamada

Functional Genomics, Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, Japan

Population structure can produce variable inflation of test statistics in case–control genome-wide association (GWA) study, and genomic control (GC) is one of the method to correct the inflation of χ^2 statistics for contingency tables of case-control independency tests. When the contingency tables have low expected values, χ^2 test is inaccurate and Fisher’s exact test should be substituted for χ^2 test. However, the GC method for Fisher’s exact test has not been indicated. We propose the application of the GC method to Fisher’s exact test, using mid-P value, in which a half of the probability of observed contingency table is included in the estimation of exact P value. The method transforms the observed mid-P values into the corresponding χ^2 values $\left(\chi^2_{\mathrm{mid-P}}\right)$, and estimates the coefficient to quantitate the variable inflation by comparing the quantile distribution of $\left(\chi^2_{\mathrm{mid-P}}\right)$

with χ^2 distribution of 1 *df*. We generated simulation case-control data sets in a range of population structures ($Fst = 0$–0.001) and sample sizes (200–4,000), and applied both GC methods. GC corrected mid-P values of Fisher's exact test achieved more accurate estimation of type I error rates for given nominal significance levels and higher statistical powers compared with GC corrected P values of χ^2 test, especially in small sample sizes (around less than 800–1,000 samples). We propose our application of the GC method to Fisher's exact test gives significant contributions in the field of GWA studies.

036: Statistics of transcription factor binding sites data sets obtained by high-throughout sequencing and chromatin IP in human genome

Yuriy Orlov, Roy Joseph, Vinsensius B. Vega, M. Huss, Neil Digby Clarke

Genome Institute of Singapore, 60 Biopolis Street #02-01 Genome, Singapore

Analysis of transcription factor binding sites is an important problem in computational genomics. Chromatin immunoprecipitation (ChIP) combined with high-throughput sequencing technologies allow us to identify transcription factor binding sites (TFBS) in genome scale. As many as 10,000–40,000 sites can be found in such experiments. Examples of such genome maps include Oct4, Sox2 and Nanog transcription factors in mouse, and p. 53, c-Myc, ER in human genome. High-throughput data sets of binding sites demands integrated computational approaches, including statistical estimates of data mapping quality, sensitivity and specificity. The first computational stage of the analysis is the accurate mapping of short sequence reads (25–35 nt) to the reference genome. Sequencing biases can skew the mapped tags toward genomic regions with higher GC content, and the presence of repeat sequences can require an adjustment to the expected chance of observing moderate peaks in the density of sequence tags due to ChIP enrichment. We have developed a statistical approach based on observed and control data to filter out false positive sequence peaks (noise) To test the significance of the peaks obtained we used computer simulations as well as and the background noise distribution for ChIP reads from control (non-specific IP) data. We have also developed statistical approaches to estimate the saturation of the TFBS maps obtained from ChIP experiments (number of false negative peaks). In a recent analysis of ChIP experiments performed in embryonic stem cells, we identified potential enhancers containing up to 10 bound transcription factors. We found by a simulation that overlaps of four or more TFBS are significant. Finally, we have analyzed the chromosomal profile of sequence tags and found correlation of TFBS with open chromatin regions and the methylation markers of active transcription.

037: The nature and heritability of the retention of heterozygosity (ROH) areas in human genome

[1]**Priit Palta**, [1]Reedik Mägi, [2]Tõnu Esko, [2]Andres Metspalu, [1]Maido Remm

[1]Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu, 23 Riia St., Tartu 51010, Estonia, [2]Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, 23 Riia St., Tartu 51010, Estonia

Studies have revealed that the human genome contains areas, where consecutive markers are all showing homozygous genotypes. It can be due to either loss of heterozygosity (LOH), when one allele is missing (for example due to deletion of one parental chromosome) or retention of heterozygosity (ROH), when both alleles are present but showing only homozygous genotypes. In this study, we have calculated the ROH areas for all HapMap Phase III populations and compared their locations and frequencies among populations. According to our results, ROH areas are similar among different populations, which indicate that there is either a biological mechanism which prevents their degradation or distinctive feature that makes them to appear in certain chromosomal areas. Also the heritability of ROH areas is studied in CEPH, Yoruba, Mexican, Afro-American and Kenya population samples. Main questions to ask would be: a) If the offspring has ROH, then does one or both of its parents are carriers of the same ROH? B) If one of the parents has ROH then how big is the chance, that the ROH has disappeared in offspring? These answers can help us to find out if the ROHs are mostly inheritable or do they frequently arise in certain chromosomal areas.

038: Genome annotation of *Anopheles gambiae* mosquito using tandem mass spectrometry-derived data

[4]**Akhilesh Pandey**, [1]Kumaran Kandasamy, [1]Dhanashree Kelkar, [1]Santosh Renuse, [1]Sutopa Banerjee, [1]Beema Shafreen, [1]Shivakumar Keerthikumar, [2]Ajeet Kumar Mohanty, [2]Ashwani Kumar, [3]Aditya Prasad Dash, [1]Pradip Acharya, [1]T. S. Keshava Prasad, [1]Nandini Patankar, [1]Raghothama Chaerkady

[1]Institute of Bioinformatics, Bangalore 560066, India, [2]National Institute of Malaria Research, Field Station, Goa 403 001, India, [3]National Institute of Malaria Research (ICMR), Delhi 110 054, India, [4]The Johns Hopkins University School of Medicine, Baltimore, MD 21205, United States of America

With the advent in genome sequencing technology, many genomes have recently been completely sequenced. However, a deeper understanding of genome organization including prediction of protein-coding genes remains a major challenge. We present proteomics as a robust complementary approach to annotate sequenced genomes. Here, we present genome annotation of Anopheles gambiae, a major sub Saharan African vector of malaria using exhaustive proteomics analysis, using mass spectrometry-derived data. We carried out a comprehensive mass spectrometry analysis of the proteome of A. gambiae mosquito, including its larval stages. The samples were homogenised and digested using trypsin and the extracted proteins fractionated using strong cation exchange chromatography. Each fraction was then subjected to liquid chromatography tandem mass spectrometry (LC-MS/MS) using a quadrupole time-of-flight mass spectrometer. The MS/MS data was searched against non-redundant protein database of all species of Anopheles, Aedes and Drosophila genera. This approach allowed us to validate a number of proteins that were labelled as 'hypothetical' in the A. gambiae databases. We were also able to identify proteins that were missed in A. gambiae protein databases but were either known or predicted in related species. An alternative approach that we took was to search the MS/MS data against a six frame translation of the genome of A. gambiae. Any peptides that were identified based on genomic sequence but were absent in protein databases were further investigated. Using this 'genome search' strategy, we identified a number of novel genes for which there was no evidence either from gene prediction programs or from transcriptomic studies. We are in the process of validating our findings using RT-PCR assays. Overall our studies show that proteomics is a good complement to transcriptomic and gene prediction approaches to annotate genomes accurately and should be used routinely.

039: In silico studies on structure and function of phylogenetically related dengue virus nonstructual and human (host) proteins

Harshwardhan Poddar, Joydip Chatterjee, Ayan Sadhukhan, Somjeet Dutta, Rudrashis Chakraborty, Anupam Dash, Bodhisattwa Saha, Abhishek Dan, Nandan Kumar Jana

Dept. of Biotechnology, Heritage Institute of Technology, Chowbaga Road, Kolkata-700107, WB, India

Dengue virus NS1 protein presents the closer phylogenetic correlation to CD61 than fibrinogen and the other two platelet integrin/adhesin relating proteins (CD41 and CD49B), (by Wiwanitkit, Int J Genomics Proteomics, 1, 2004). Similarly, the present study has been carried out to search for phylogenetic correlations between other Nonstructural proteins of Dengue virus and the respective host proteins. For these Nonstructural Proteins NS2A, NS2B, NS3, NS4A, NS4B and NS5 were chosen (from Dengue virus type-1, 2, 3 and 4) to check their correlations with human host proteins Interleukin 12A, Interleukin 12B, Interleukin 8, Interleukin 10, Integrin-alpha 8, Integrin-beta 5, Integrin-beta 7, Human RANTES, Human Interferons (IFN-alpha, beta, gamma), Human STAT2 protein and Tumor Necrosis factor (TNF ligand superfamily number 9 and 10). Our objectives were in silico studies of phylogenetically close proteins focusing on—secondary and tertiary structures based on homology modeling, domains and motifs related to function of these proteins and protein disorders; to predict relationships between observed structure and function at any level of phylogenetically close Dengue viral and human host proteins using bioinformatics tools like ClustalW, 3D Jigsaw, Prosite, ProFunc, PDBsum, etc. This is solely a review work. The predicted structures and functions, if reviewed further, may lead to important insights into the mechanism of Dengue viral pathogenesis. The data presented here may also have certain archival value, while defining targets for Dengue disease therapeutics in general and lead to understanding of the role of these proteins in the disease pathway in particular.

040: SSPred: a prediction server based on SVM for the identification and classification of proteins involved in bacterial secretion systems

Sachin Pundhir, Anil Kumar

School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore (MP), India

Protein secretion systems used by almost all bacteria are highly significant for the normal existence and interaction of bacteria with their host. The accumulation of genome sequence data in past few years has provided great insights into the distribution and function of these secretion systems. In this study, a support vector machine (SVM)-based method, SSPred has been developed for automated functional prediction of proteins involved in secretion systems and classifying them into five major secretion systems (Type-I, Type-II, Type-III, Type-IV and Sec systems). The dataset used in this study for training and testing was obtained from KEGG and SwissProt database. The dataset was curated in order to avoid redundancy. To overcome the problem of imbalance in positive and negative dataset, ensemble of SVM modules each trained on a balanced subset of the training data were used. The SVM modules were trained and optimized with different protein sequence features like amino-acid composition, dipeptide composition and physico-chemical properties. The SVM module based on dipeptide composition classified secretion system proteins and non-secretion system proteins with an accuracy of 83.87% when evaluated using fivefold cross-validation. It is marginally better than the modules based on amino-acid composition (Accuracy = 83.56%) and physico-chemical properties (Accuracy = 80.96%). The method based on dipeptide composition is further able to predict five major classes of secretion systems with an overall accuracy of 81.56% and Matthew's correlation coefficient (MCC) of 0.64. In order to improve the accuracy, we developed a hybrid module using all the features of a protein, resulting in an input vector of 456 dimensions (20 amino acid compositions, 400 dipeptide compositions and 36 physico-chemical properties). The hybrid module was able to achieve an accuracy and MCC of 85.09% and 0.70, respectively in distinguishing proteins involved in secretion systems from others. Similarly, the overall accuracy of hybrid modules in predicting five major classes of secretion systems was 82.83% with MCC of 0.66. For fair evaluation of the hybrid modules, their performance was tested on independent/blind dataset. The method correctly predicted 85.54% proteins as being involved in secretion systems with MCC of 0.72. For five major classes, the overall accuracy was 84.52%. SSPred is available as a freely available World Wide Web server—SSPred at http://www.davvbiotech.res.in/SSPred.

041: Structural domain prediction for transcription factor binding sequences

Padmavathi Putta, Chanchal K. Mitra

University of Hyderabad, P.O. Central University, Hyderabad 500 046, India

Transcription factors are proteins that play crucial role in transcription and in the regulation of gene expression. Transcription factors are classified into several types: zinc-finger, helix-loop-helix, leucine zipper, helix-turn-helix, etc. depending on their structure of the DNA binding domains. One of the remarkable aspects and important step of the transcription is the sequence specific DNA binding selectivity exhibited by most of the transcription factors. They are capable of selecting correct binding sequences in the genome out of vast number of potential alternative sites. These DNA sequences may be of around 8–10 nucleotides in size. Recognition of these sequences upstream to the promoter or enhancer regions requires a minimum of two or three transcription factors and can be mediated by the miRNA associated with the transcription factors. Once the transcription factor recognizes and binds to the DNA, it activates other transcriptional factors and the transcriptional machinery to facilitate transcription. We have downloaded 1871 human promoter sequences from SIB-EPD and studied the sequences from -100 to +100 with respect to transcription start site. We have looked at the upstream regions of the various promoter regions and searched for common subsequences using a conventional program. The eukaryotic promoter region is usually complex and do not have conserved sequences as seen for prokaryotes. We have identified the distribution of most common 50 hexanucleotide sequences that are rich in GC content on both promoter side and the coding side and observed for their frequency of distribution. The sigmoidal behavior of the distribution of these hexanuclotides suggests that there exists some internal cooperative manner in these sequences on both upstream and downstream regions of TSS. In continuation of the above studies we have searched the JASPAR database and extracted the binding profiles for all the structural classes of transcription factors. We have compared our previous results with the binding profiles and evaluate the matrices for further analysis to identify the frequency distribution of subsequences that are located upstream of the promoters. From these results we can identify the most predominant subsequences that can be recognized by the transcription factors. These results provide basic information that can be implemented for further classification of

transcription factors based on their binding selectivity to the subsequences in the DNA template.

042: NetPath: a public resource of curated signal transduction pathways: development of an initial set of immune signaling pathways

[1]**Rajesh Raju**, [1]S. Sujatha Mohan, [1,3]Balamurugan Periaswamy, [1,3]Suresh Mathivanan, [1]Kumaran Kandasamy, [1]Shivakumar Keerthikumar, [1]Shubha Suresh, [1]Anuradha Nalli, [1]Bincy Jacob, [1]Deepthi Pantula, [1]Salil K. Sukumaran, [1]Sapna Upendran, [1]Sashi Kanth Gollapudi, [1]Shweta Gupta, [1]Tanima De, [1]T. S. Keshava Prasad, [1]G. S. Sameer Kumar, [1]Malabika Sarker, [1]Sudhir Gopal Tattikota, [2,3]Gary D. Bader, [3]Chris Sander, [1,4]Akhilesh Pandey

[1]Institute of Bioinformatics, International Tech Park, Bangalore-560066, India, [2]University of Toronto, Toronto, Ontario M5S 3E1, Canada, [3]Memorial-Sloan-Kettering Cancer Center, New York 10021, United States of America, [4]The Johns Hopkins University School of Medicine, Baltimore, MD 21205, United States of America

Existing pathway resources in the public domain are mostly limited to specific lineage of interactions to define only a set of sub-pathways. NetPath is a freely available resource that contains comprehensive information on signal transduction pathways. The curation of pathway information and storage of data is achieved through a tool we developed called PathBuilder. PathBuilder facilitates easy retrieval and export into community standardized data structures such as BioPAX, PSI-MI and SBML. As an initial step, we have generated a set of immune signaling pathways namely B cell receptor, T cell receptor and 8 Interleukin pathways. Each pathway has information pertaining to protein-protein interactions, enzyme catalysis events and translocation of proteins within and across the sub-cellular compartments under ligand stimulation. The protein-protein interactions include binary interactions and interaction with functional complexes. For each catalysis or modification events that influences the status of molecules, the upstream enzyme and site of modification is curated whenever it was available. This resource also contains genes which are reported to be differentially expressed under each pathway obtained by both microarray and non-microarray experiments from human source. A novel statistical method, which was recently reported by Draghici et al. (2007), was used to measure the impact factor for pathways in gene expression experiments. This approach can capture the magnitude of the expression changes of each gene, the position of the differentially expressed genes on the given pathways, the topology of the pathway that describes how these genes interact and the type of signaling interactions between them. The present set of immune signalling pathways include more than 1,100 reactions annotated from the literature and $>$ 2,400 instances of transcriptionally regulated genes in response to receptor stimulation from over 4,400 published articles. NetPath is available as a worldwide web resource from http://www.netpath.org/.

043: MalVac: database of malarial vaccine candidates

Ramu Srinivasan Ramachandran, Rupanjali Chaudhuri, Faraz Alam Ansari, Shakil Ahmed, Harinder Vir Singh

Institute of Genomics and Integrative Biology, Mall Road, Delhi-110 007, India

Background: Malaria is a major killer disease. Annually more than 500 million cases are reported and over 1 million deaths occur. The sequencing of genomes of the Plasmodium species causing malaria, offers immense opportunities to aid in the development of new therapeutics and vaccine candidates through Bioinformatics tools and resources.
Methods: The starting point of MalVac database is the collection of known vaccine candidates and a set of predicted vaccine candidates identified from the whole proteome sequences of Plasmodium species provided by PlasmoDb 5.4 release(31st October 2007). These predicted vaccine candidates are the adhesins and adhesin-like proteins from Plasmodium species, *P. falciparum*, *P. vivax* and *P. yoelii*. Subsequently, these protein sequences were analysed through 20 publicly available algorithms to obtain Orthologs, Paralogs, Beta-Wraps, TargetP, TMHMM, SignalP, CDDSearch, BLAST with Human Ref. Proteins, T-cell epitopes, B-cell epitopes, Discotopes, and allergen predictions. All these information were collected and organized with the ORFids of the protein sequences as primary keys. This information is relevant from the view point of Reverse Vaccinology in facilitating decision making on the most probable choice for vaccine strategy.
Results: Detailed information on the patterning of the epitopes and other motifs of importance from the viewpoint of reverse vaccinology has been obtained on the most probable protein candidates for vaccine investigation from three major malarial species *P. falciparum*, *P. vivax* and *P. yoelii*. Analysis data are available on 161 adhesin proteins from *P. falciparum*, 137 adhesin proteins from *P. vivax* and 34 adhesin proteins from *P. yoelii*. The results are displayed in convenient tabular format and a facility to export the entire data has been provided. The MalVac database is a 'community resource'. In this spirit, we encourage users to export data and further contribute by value addition. Value added data may be sent back to the community either through MalVac or PlasmoDB.
Conclusion: A web server MalVac for facilitation of the identification of probable vaccine candidates has been developed. Availability: The MalVac server can be accessed at http://malvac.igib.res.in/
.

044: Evolution of diversity in Polycomb homologues

Senthilkumar Ramamoorthy, Rakesh Mishra

Center for Cellular and Molecular Biology, Uppal Road, Hyderabad 500 007, India

Polycomb group (PcG) proteins maintain cell type specific expression pattern of genes by altering chromatin organization. These proteins function as multi protein complexes, viz., PRC1 and PRC2. The PRC2 recognizes the Polycomb Response Elements (PREs) in the genome and tri methylates H3K27. Subsequently, PRC1 binds to these methylation marks and maintains the locus in repressive state. The chromodomain of PC which is a component of PRC1 recognizes the H3K27 trimethylation mark. Polycomb was first discovered in Drosophila and while insects have one Pc gene, most vertebrates have five Pc homologues, which are called Cbx2, Cbx4, Cbx6, Cbx7 and Cbx8. To understand the importance of having more homologues and their possible functional specificity, we carried out bioinformatics analysis of vertebrate Pc homologues. We identified 85 Pc homologs from different vertebrate species from protein and genome databases for this purpose. All homologues have N-terminal chromodomain and C-terminal Polycomb Repressor (PCR) box. We discovered conserved domains specific to each of five Cbx members. We also found a novel insect specific conserved motif by comparing Pc homologues of this group of animals. Presence of unique and conserved motifs in each Cbx proteins indicates that early during vertebrate evolution Pc gene expanded into several copies and each one of them evolved independently by acquiring functional specificity. Interestingly, expansion of Pc homologues in vertebrates also

coincides with the increased number of hox complexes in vertebrates. So much so that in teleosts that have more hox complexes than other vertebrates also have more homologeus of Pc gene. The maintenance of distinct Pc homologues indicates their distinct roles in regulation of genes, particularly the Hox genes. Our findings open up the way to investigate the molecular basis of gene expansion and maintenance by acquisition of novel motifs during the evolution of complexity in vertebrates.

045: Genome wide application of DNA stability for the annotation of promoter regions

Vetriselvi Rangannan, Manju Bansal

Indian Institute of science, Bangalore, India

Identification and annotation of promoter regions in genomes has drawn significant attention in structural and functional genomics community, due to their important role in controlling gene expression. Prokaryotic as well as eukaryotic promoter sequences exhibit several common sequence and structural features, but very little sequence motif consensus is seen. However, it has been proved that promoter sequences possess certain sequence dependent structural properties, such as lower stability, higher curvature and less bendability, when compared with their neighboring regions. Based on the difference in stability between neighboring upstream and downstream regions in the vicinity of experimentally determined transcription start sites, a promoter prediction algorithm has been developed to identify prokaryotic promoter sequences in whole genomes. Our results prove that while the promoter regions are overall less stable, their average free energy varies compared to other genomic regions, depending on the GC composition of the genomic sequence. In this method the average free energy (E) over known promoter sequences and the difference (D) between E and the average free energy over downstream random sequences (REav) are used to search for promoters in the genomic sequences. Thresholds (E and D) have been generalised for every 5%GC interval. A protocol that has been successfully applied for microbial promoter identification and their prediction on a large-scale to whole genomic sequences will be presented. High reliability of 70 and 61% for *E. coli* and *B. subtilis*, respectively, was achieved on carrying out whole genome annotation for proximal promoters in these genomes.

046: Motifs as markers in diseased proteins: in silico investigations of potential drug targets

N. Rathankar Rao, S. Sandeep Kavilu kudige, P. Sandesh Kanchugal, G. Nagendra Holenarasipur

Dept. of Biotechnology, Sir M. Visvesvaraya Institute of Technology, Near Hunsamaranahalli, Via Yelahanka, Bangalore 562 157, India

Proteins contain structural moieties characterized by domains, motifs and folds that are often directly involved in its function. It is established that, when these motifs become malfunctional or undergo mutation, onset of diseases occurs. Hence, investigations to establish a plausible relationship between disorders/diseases and the signature of motifs that may be involved, have being carried out in this work. The aim was thus to find the role of motifs in diseased proteins, characterize them as possible markers towards diagnosis, model them as potential drug targets and propose the development of a common drug to tackle family of disorders.

As a case study, a total of 120 protein sequences for four neuro degenerative disorders (Alzheimers, Parkinson, Huntington, prion disease) were extracted using the KEGG pathway. Every sequence in the diseased pathway was then subjected to a database search, which resulted in 409 homologous sequences with percentage identity greater than 25%. These 409 sequences were then subjected to pattern search using PRATT tool that yielded 510 non-overlapping motifs. These motifs were nomenclatured as gene-name_MXX (where gene-name as per KEGG database, M is the motif and XX for an integer), to facilitate proper identification. Validations of these motifs were carried out against the SWISSPROT database using the SCANPROSITE tool to infer whether these motifs were implicated in diseases or not. Interestingly, 289 unique motifs implicated in several diseases, including non-neurological disorders were obtained. An interaction map was generated to decipher the probable relationships between these 289 motifs.

Interestingly, 8 motifs related to Alzheimer and Parkinson's, 3 related to Parkinson with other non-neurological disorders and 9 with Alzheimer and other non-neurodegenerative diseases were identified. Two motifs, one belonging to Alzheimer(A2M_M01) and another to Prion disease (LAMC1_M01), were also found to be occurring in seven other non-neurological diseases. The results highlighted that sequences common to Parkinson, Alzheimer's and Bone disease, contained 2 characteristic motifs (GPR37_M02 and GPR37_M04), indicating that these motifs are signatures of neurological and non-neurological diseases as well. Thus, this work suggests that motifs could be used as potential markers for diagnosis and drug targets for diseases containing such common signatures.

047: ELAN: a server based tool for genome wide analysis of mobile genetic elements

[1,5]**Kamal Rawal**, [2]Sudha Bhattacharya, [1,3]Alok Bhattacharya, [1,4]Ram Ramaswamy

[1]School of Information Technology, Jawaharlal Nehru University, New Delhi 110 067, India, [2]School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110 067, India, [3]School of Life Sciences, Jawaharlal Nehru University, New Delhi 110 067, India, [4]School of Physical Sciences, Jawaharlal Nehru University, New Delhi 110 067, India, [5]Department of Biotechnology, JIIT University, Noida, India

Mobile genetic elements occupy significant proportion of eukaryotic genomes. They are involved in number of important cellular functions. ELAN is a suite of tools for genome wide analysis of mobile genetic elements. It finds distribution and nature of mobile genetic elements. DNA SCANNER is a part of ELAN which analyses insertion sites of mobile genetic elements for the presence of various physicochemical signals. Insertion Site Finder (ISF) is a machine learning tool which incorporates information derived from DNASCANNER and uses support vector machines to classify DNA sequences into insertion sites and non insertion site classes. ELAN has been applied to wide variety of organisms. It has identified distributions of several mobile elements such as Alu in various organisms such as Human, Mouse, Drosophila, E. histolytica etc. DNA SCANNER has identified common set of statistically important signals flanking insertion sites in various genomes suggesting common insertion mechanism operating in wide variety of organisms. ISF has emerged to be an important tool for insertion site prediction as it has shown high accuracy levels (65–90%). The dataset and information derived during analysis will serve as bench marking resource in future for various analyses. Large data has been organized into web

portal as well as relational database named as InSiDe which is available online at http://nldsps.jnu.ac.in/bioit/ccbb/elan.html.

048: Identification of translational regulatory elements in 5′ UTRS of HIV genomes: possible targets for designed Aptamers

Paushali Roy, Sayak Ganguli, Priyanka Dhar, Abhijit Datta

Bioinformatics Infrastructure Facility, Presidency College, 86/1 College Street, Kolkata, 70073, India

UTRs of mature MRNA are known to play crucial roles in the post-transcriptional regulation of gene expression, subcellular localization and stability. Delayed translational silencing of ceruloplasmin (CP) mRNA by gamma-interferon (IFN-gamma) is directed by a structural element in its 3′ and 5′ untranslated region. Translational silencing requires the binding of a cytosolic inhibitor to the CP 3′UTR and all essential elements of mRNA circularization, i.e. eukaryotic initiation factor 4G, poly-A binding protein and poly-A tail. The 29 nt structural element, denoted as IFN-gamma-activated inhibitor of translation (GAIT), consisting of a 5 nt terminal loop, a weak 3 bp helix, an asymmetric internal bulge and a proximal 6 bp helix, has been experimentally demonstrated to be sufficient for translational silencing both in vitro and in vivo in human. Such GAIT elements are also found to be present within the 5′ untranslated regions (UTRs) of Simian Human Immunodeficiency Virus and are 64 in number. The identification of such regulatory elements in the viral genome suggests that if engineered aptamers can be designed resembling these sequences then these could probably aid to regulate the expression of HIV and thus bring an end to all misery of the human race.

049: Computational identification of functional SNPs at 5′ splice sites

[1]**Ravi Sachidanandam**, [2]Adrian R Krainer, [2]Xavier Roca

[1]Dept. of Genetics and Genomic Sciences, Mount Sinai school of medicine, 1425 Madison Avenue, New York, NY 10029, United States of America, [2]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, United States of America

Only a small fraction of the millions of SNPs known to exist in the human genome have a direct effect on the expression of genes. Large scale screens using association studies are aimed at ultimately identifying these functional SNPs, which may often be involved in complex diseases. We identify functional SNPs that occur in 5′ splice sites (5′ss) and allow for a direct probing of the role of their host genes. We do this by devising new measures of 5′ss strength. The commonly used measure of 5′ss efficiency, position weight matrices, sometime fail to predict 5′ss disrupting SNPs that can cause diseases. Among the known diseases caused by 5′ss mutations are Fanconi anemia, hemophilia B, neurofibromatosis, and phenylketonuria. By using comparative genomics, we identify pairwise dependencies between 5′ss nucleotides as a conserved feature of the entire set of 5′ss. These dependencies are also conserved in human-mouse pairs of orthologous 5′ss. Many disease-associated 5′ss mutations disrupt these dependencies, as can some human SNPs that appear to alter splicing. The consistency of the evidence across a variety of sources signifies the relevance of this approach. We use the pairwise dependencies identified above to identify 5′ss SNPs that may play a role in complex diseases. We have identified new measures of 5′ splice site strength and identified functional human SNPs in 5' splice sites that probably have a role in complex diseases.

050: From genome wide association towards underlying genetic association networks: GWANA

[1,2]**Juha Saharinen**, [3]Heikki Mannila, [4,5,6]Aarno Palotie, [1,5,6]Leena Peltonen

[1]Department of Molecular Medicine, National Public Health Institute of Finland, Biomedicum Helsinki, Finland, [2]Genome Informatics Unit, Biomedicum Helsinki, Finland, [3]HIIT Basic Research Unit, Department of Computer Science, University of Helsinki, Helsinki, Finland, [4]Finnish Genome Center, University of Helsinki, Biomedicum Helsinki, Finland, [5]Wellcome Trust Sanger Institute, Cambridge, United Kingdom, [6]The Broad Institute, MIT, Boston, MA, United States of America

The genetic component in complex diseases is acknowledged to consist of a combinatorial network of multiple susceptibility genes. Despite the recent avalanche of novel genes identified by genome wide association studies, the analyses of joint effects of genes have to some extent lacked tools tackling all the association signals in a comprehensive way to decipher the pathways behind complex traits.

We present here an analytical approach of discovering the genetic association networks involved in complex diseases in case-control settings. Our Genome Wide Association Network Analysis (GWANA) method aims to captivate comprehensively the associated genetic components behind complex traits, including loci representing "grey zone" signals, not reaching the genome wide significance for individual variants. GWANA combines the genome wide association analysis with sensitive pathway analysis as well as analysis of protein-protein interaction networks, therefore detecting the altered genetic association networks. Further, GWANA uses rigorous permutations for validating the results and is able to work with multiple sources of biological pathways/networks.

GWANA was applied to analyze 12 case-control GWA datasets in 10 different complex diseases, namely autism, bipolar disorder, Chron's disease, coronary-heart disease, hypertension, multiple sclerosis, rheumatoid arthritis, types 1 and 2 diabetes and Schizophrenia. The method proved to identify biologically relevant genetic association networks, likely to be involved in the etiology of these diseases.

051: The Sri Lankan genome variation database

P. S. Samarakoon, R. W. Jayasekara, V. H. W. Dissanayake

Human Genetics Unit, Faculty of Medicine, University of Colombo, Kynsey Road, Colombo 8, Sri Lanka, Sri Lanka

Sri Lankan Genome Variation Database (SLGVD) is a database of genetic variations found in Sinhalese, Sri Lankan Tamils and Moors, the three major ethnic groups in the Sri Lankan population. Studies of variations in genes among different groups of individuals in the Sri Lankan population have grown rapidly during last few years. These studies generated large amount of genetic data which is important to study the occurrences of diseases that differ across ethnic groups. There is therefore a need for a central repository of this data. The SLGVD was created to fulfill this void. This offers a web based access to genetic variation information of Sri Lankan people. It would also be an important informatics tool for both research and clinical purposes to retrieve and deposit human variation data. The database was designed confirming to guidelines issued by the Human Genome Variation Society (HGVS). The variation data cataloged in SLGVD were derived

from research performed by Sri Lankan Scientists. Addition to variation data each variation links with the relevant entries of Online Mendelian Inheritance in Man (OMIM), SNP and Genbank databases at National Center of Biotechnology Institute (NCBI). For each variation, genotype and allele frequencies of different ethnic groups are represented in numerical and graphical format. SLGVD can be publicly accessed from http://hgucolombo.org/default.aspx.

052: dbSNP-lite: a derivative from each build of dbSNP that tracks and consolidates core rsID marker and allele changes from each previous build

[1]**Pallavi Sarmah**, [2]Gudmundur A. Thorisson, [1]Debasis Dash, [2]Anthony J. Brookes

[1]G.N. Ramachandran Knowledge Center for Genome Informatics, Institute of Genomics and Integrative Biology (CSIR), Mall Road, Delhi, India, [2]University of Leicester, Leicester, United Kingdom

The content of dbSNP provides a useful basal layer of human DNA variation that other databases may wish to use as a framework to hang various annotations upon. The Human Genome Variation Database of Genotype to Phenotype associations (HGVbaseG2P: http://www.hgvbaseg2p.org/) uses dbSNP records precisely this way, to underpin extensive summary-level genetic association data. However, problems can arise when there are marker deletions or mergers, or allele deletions or substitutions, upon the release of a new dbSNP build, since other important database information may have been connected to those now altered marker details. To overcome this problem for HGVbaseG2P, we have devised a means for tracking all such changes and allowing for them by altering key data relationships. The system—called dbSNP-lite—not only makes suitable allowances for all marker and allele changes, but also executes a comprehensive check for any other rsID differences between old and new builds of dbSNP. At the heart of dbSNP-lite lies an automated dbSNP parser that integrates the data of a new build of dbSNP with that of HGVbaseG2P. The processing logic handles each dbSNP marker in turn, and involves a decision tree that checks rsID presence/absence, rsID merge/deletion history, strand representations, flank sequence equivalences, allele identities, strand to allele consistency, and chromosome map location(s). The parser also identifies any UnDeleted and UnMerged Markers. The output is an XML structured file that represents the validated new content of dbSNP, plus the specifics of any errors identified. After manual curation of any gross errors that cannot be allowed for automatically, the final dataset is passed into HGVbaseG2P to update the marker content to match the latest dbSNP build, at the same time suitably adjusting all connections to other information in the database and keeping a detailed log of all build changes identified. Thus, the database is kept fully up to date with the latest dSNP build, without any loss or corruption of marker/allele to phenotype associations, or allele/genotype frequency information. This open source system can be further extended and adapted to fulfill a similar role for any other database that has a similar dependency upon dbSNP. [The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 200754—the GEN2PHEN project.]

053: Multiple pseudoknot motifs in Wilms' Tumor associated genes: an in silico perspective

Sutapa Sengupta, Sayak Ganguli, Abhijit Datta

Bioinformatics Infrastructure Facility, Presidency College, 86/1 College Street, Kolkata, 700073, India

Wilms' tumour (WT; nephroblastoma), a kidney neoplasm, is one of the most frequently occurring solid tumours of childhood. It arises from the developing kidney by genetic and epigenetic changes that lead to the abnormal proliferation of renal stem cells (metanephric blastema). WT serves as a paradigm for understanding the relationship between loss of developmental control and gain of tumourigenic potential. In particular, loss of function of tumour suppressor genes has been implicated in the development of WT, and the Wilms' tumour suppressor gene WT1 (at chromosome 11p13) was the second tumour suppressor gene to be cloned, after the retinoblastoma gene RB1. WT1 plays an essential role in kidney development, but is mutated in only approximately 20% of WTs, which suggests that further lesions and genetic loci are involved in Wilms' tumourigenesis. Other chromosomal regions associated with WT include 7p, 11p15, 16q and 17q. Although many of these loci probably contain tumour suppressor genes, imprinted genes (genes showing expression of only one parental allele) and oncogenes have also been implicated in WT. The in silico analysis focused on the RNA analysis of the genes associated with the disease. We have located multiple pseudoknot motifs in the genes as well as numerous catalytic motifs. The identification of pseudoknot motifs in the eukaryotic gene set is a breakthrough discovery since it further establishes the evidences of horizontal gene transfer. The pseudoknot motifs undertake diverse functions within the living system. These functions include forming the catalytic core of various ribozymes, self-splicing introns, and telomerase. They also induce the frame shifting of ribosomes using the help of a slippery sequence. These properties make them attractive drug targets for treatment of this deadly disease in the future.

054: Fully disordered proteins in human proteome

Gajinder Pal Singh, Debasis Dash

Institute of Genomics and Integrative Bioloy, Mall Road, New Delhi, India

Proteins with long structurally disordered regions are very common in eukaryotes. How structurally disordered regions perform their functions is a subject of intensive current research. An even more interesting class of proteins is proteins which are almost fully disordered, because any functions that these proteins do have to be mediated by disordered regions. We find that eukaryotes and particularly human have significant number (thousands) of proteins predicted to be fully disordered. A majority of these proteins in human are functionally uncharacterized, but other functions overrepresented in these proteins include transcription factors and enzyme inhibitor activities. Most of these proteins do not have a homolog in mouse suggesting rapid evolution of these proteins. We believe experimentally studying the role of these fully disordered proteins can prove to be very fruitful in understanding how lack of structure relates to function.

055: DLact: an antimicrobial resistance gene database

Reema Singh, Suchir Arora, Harpreet Singh

Indian Council of Medical Research, V. Ramalingaswami Bhawan, Ansari Nagar, New Delhi-110029, India

Lactamase genes are the main cause of antimicrobial resistance in the bacterial pathogens. Here, we introduce DLact, a curated collection of lactamase like proteins. The availability of genomic data provided us an opportunity to develop such a database. Using the completely sequenced bacterial chromosomes and plasmids we identified putative antimicrobial resistance genes and have developed a database. The growing database currently contains 2021 lactamase like genes from

814 sequenced bacterial genomes. Of which 1972 (97.57%) were present on chromosome and only 49 (2.42%) were present on plasmids. Database content can be searched by using text and sequence queries. Diversity at the taxonomic, microbial ecology and domain length was studied. DLact database may be used to develop diagnostic primers and probes and in identifying pathways controlling/affecting the expression of antimicrobial genes. As the curated collection of sequences available for a protein-encoding gene, DLact provides a resource for researchers interested in comparative protein modeling and drug designing,as well as those interested in broad subjects such as lateral gene transfer and Codon usage. We have also created in-house perl scripts for regular updation of the database. To share this resource with the scientific community, we have designed and implemented a web interface for DLact by using open source tools and it is available at http://ijmr.in/Dlact/.

056: Understanding human genome organization based on physico-chemical properties of DNA

P. Singhal, G. Khandelwal, S. Tripathi, B. Jayaram

Department of Chemistry and Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology, Hauz Khas, New Delhi, 110 016, India

An ab initio model for gene prediction is proposed based on physico-chemical characteristics of codons calculated from molecular dynamics (MD) simulations. The model requires a specification of three calculated quantities for each codon: the double helical trinucleotide base pairing energy, the base pair stacking energy and an index of the propensity of a codon for protein-nucleic acid interactions. The base pairing and stacking energies for each codon are obtained from recently reported MD simulations on all unique tetranucleotide steps Beveridge et al. (2004), and the third parameter is assigned based on the conjugate rule previously proposed to account for the wobble hypothesis with respect to degeneracies in the genetic code Jayaram (1997)[2]. The third interaction propensity parameter values correlate well with MD calculated solvation energies and flexibility of codon sequences as well as codon usage in genes and amino acid composition frequencies in $\sim$ 175000 protein sequences in the Swissprot database. Assignment of these three parameters for each codon enables the calculation of the magnitude and orientation of a cumulative three dimensional vector for any given DNA sequence of any length in each of the six genomic reading frames. Analysis of 372 prokaryotic genomes comprising $\sim$ 350000 genes shows that the orientations of the gene and non-gene vectors are well differentiated and make a clear distinction feasible between genic and non-genic sequences at a level equivalent to or better than currently available knowledge based models trained on the basis of empirical data, presenting a strong support for the possibility of a unique and useful physicochemical characterization of DNA sequences from codons to genomes (Singhal et al. 2008; Dutta et al. 2006).

We present here an adaptation of the above described physico-chemical approach for the analysis of all the 24 chromosomes of Human Genome. Results showed a strikingly similar orientation of the genes of all the chromosomes pointing to the promise of developing a physcio-chemcial understading of human genome organization.

References

Beveridge DL et al (2004) Biophys J 87:3799–3813
Jayaram (1997) J Mol Evol 45:704–705
Singhal et al (2008) Biophys J 94(11):4173–4183
Dutta et al (2006) J Chem Inf Model 46(1):78–85

057: A Novel Glocal alignment of protein sequences with evolutionarily conserved subsequences

[1]**Balasubramanian Sriram**, [2]R. Sowdhamini

[1]Indian Institute of Technology, Kanpur, India, [2]National Center for Biological Sciences, Bangalore, India

Introduction: A novel Global Local Alignment Algorithm addressing the problem of aligning two protein sequences globally while conserving evolutionarily unchanging, smaller subsequences occurring in one of the sequence, is proposed. Within these subsequences, insertions and deletions are restricted implying evolutionary conservation of structure. Thus the alignment is of global character with local restrictions about these constrained subsequences and hence glocal. The algorithm is implemented with the methods dynamic programming and iteration and tested successfully on small sequences. A biological interpretation with an evolutionary basis, of the logic of the algorithm is provided including a step by step analysis of the algorithm from a biological point of view of protein alignment. An intensive run time analysis of the algorithm is also performed. Problem Formulation: Align two sequences S1 :- a1…ap and S2 :- b1… bq, conserving n subsequnces in S1 SS1 :- ai[1], ai[1] + 1 … ai[1] + j[1] …SSn :- ai[n],ai[n] + 1 … ai[n] + j[n], i[p] indicates the start of the pth subsequence and j[p] indicates it's length. Also ai[p] > ai[p-1] + j[p-1]for all p. Algorithm: The Glocal algorithm, ALGX, works best in the limit that the conserved sequences are smaller in size (approx 1/10th the sequence size) and in number. (1) NEEDLEMAN WUNSCH global alignment algorithm is applied to S1 and S2. If all the subsequences are internally conserved i.e. without indels after this step, then algorithm is complete. (2) If gaps are present between the residues of the conserved subsequence (i.e. SSi, i from (1 − n)) in the aligned sequence, BRUTE ALGN routine is applied within each subsequence for each such subsequence. 2(a). BRUTE ALGN routine:- The non contagious residues of the conserved segment are converted to a contiguous sequence block displacing the internal gaps to one side (right). Iteratively, alignment score is computed for each position of the conserved block, aligned with the residues of the other sequence (which is aligned globally with this sequence in step 1), till all displaced internal gaps on the other side (left). The position having the maximum alignment score is fixed as position of subsequence (with the internal gaps after step now occurring on either side). (3) Needleman Wunsch is applied individually to each set of residues on the either side of the conserved regions with the corresponding residues on the aligned sequence S2. This gives the Glocally aligned sequences.

058: Co-transcription of genes into single transcripts: another regulatory mechanism for gene expression in vertebrates

Tulika Prakash Srivastava, Vineet K. Sharma, Naveen Kumar, Tadayuki Takeda, Ritsuko Ozawa, Maki Mushiake, Reina Okumura, Yuichiro Nishida, Takayoshi Fujikake, Todd D. Taylor

MetaSystems Research Team, Computational Systems Biology Research Group, RIKEN Genomic Sciences Research Complex (GSC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045, Japan

Co-transcription of two distinct genes (child genes) into a single transcript (conjoined gene) has not been well explored. Either it is a somewhat rare phenomenon, or the current methods of genome annotation are not sensitive enough to identify such genes. Towards this we have designed a new computational algorithm 'Conjoin' for the identification of conjoined genes in any genome given its messenger RNA

or EST information. Applying Conjoin to the human genome, we have identified nearly 750 conjoined genes of variable lengths, some with multiple isoforms. We have so far confirmed the existence of more than 200 of these conjoined genes using RT-PCR and sequencing. In view of the fact that we observed several cases of conjoined genes occurring in the human genome, it appears that these conjoined transcripts are arising out of novel functional requirements and are not merely artifacts of transcription. However, the underlying mechanism controlling the formation of such conjoined genes in human and other vertebrate genomes remains to be explored. In order to confirm the presence of conjoined genes in other vertebrates, we implemented the Conjoin algorithm on both the mouse and chimpanzee genomes. It is remarkable to observe that the number of conjoined genes in mouse is far less than that in human, even though there is roughly the same amount of mRNA/EST data available. Thus it appears that the conjoined genes might be performing some novel functions and are contributing to human complexity as compared to other lower organisms. Therefore, we carried out a detailed functional analysis of the human conjoined and participating child genes. Further, in order to explore the intrinsic mechanisms for the formation of conjoined genes, the 5′ and 3′ flanking regions of the child genes were analyzed to search for the presence of any alternate or common regulatory elements that might be controlling the formation of conjoined genes. Ten regions in the human genome were selected which satisfy the minimum requirement for the formation of a conjoined gene as observed by our analysis and where no prior evidence (mRNA/EST) for the existence of such genes is available. In eight (80%) of these selected regions existence of a conjoined gene could be confirmed using RT-PCR and sequencing. Finally a comprehensive database of all the human conjoined genes is designed to provide a repository of these specialized genes with detailed information about each gene.

059: Identifying components of Alzheimer's disease using a genome-wide Endoribonuclease-prepared siRNAs library

Vineeth Surendranath, Jitin Bali, Frank Buchholz, Kai Simons, Bianca Habermann, Lawrence Rajendran

Max Planck Institute of Molecular Cell Biology and Genetics/ Scionics Computer Innovation GmbH, Pfotenhauer Str 108, 01307 Dresden, Germany

Sequential processing of amyloid precursor protein (APP) by β- and γ-secretase results in the production of the amyloid-β (Aβ) peptide, that is causally linked to Alzheimer's Disease. Several mechanisms control APP processing but the extent to which this process is regulated by cellular signaling machinery is still not completely understood. We performed a genome-wide analysis of human kinases and phosphatases using small interfering RNAs combined with network analysis to identify key regulators of APP processing. In order to characterize the hits, instead of selecting one or more favourite genes of choice, we used network analysis to connect all the hits from the screen to key players in APP processing. Towards this end, we looked to see how the hits from the screen and the already established genes involved in Alzheimer's disease associated in the human interactome. Computational analysis using an iterative shortest paths approach yielded a large number of possibly additional genes involved in the disease necessary to establish connections between the hits and the established genes. The subsequent question addressed was which of these genes i.e., their protein products, were specific to the network subspace defined by the established Alzheimer's disease genes and the hits resulting from the screen. This analysis yielded a set of 61 specific genes which could play a role in the associations between the hits from the genome wide screen and the already established Alzheimer's disease genes, also expanding on the potential list of genes involved in Alzheimer's disease pathogenesis. One of these connecting proteins is CDK5 which had been suggested to be involved in Alzheimer's disease. Here, we present evidence that cyclin dependent kinase-5 (CDK5) is a key regulator in Aβ production by regulating the expression of β-secretase. Our results reveal the importance of signaling pathways and protein networks in Aβ metabolism providing key insights into drug discovery and basic understanding of the disease.

060: Identification of telomeric and sub-telomeric signature sequences of small eukaryotic chromosomes using information theory

D. R. Swati

Depertment of Physics, MMV, Banaras Hindu University, Varanas-221005, U.P., India

Genome variability or plasticity is the basis for local adaptation, antigenic variation and drug resistance of pathogens like Plasmodium falciparum, Leishmania major, Trypanposoma brucei and Trypanosoma cruzi, that cause cerebral malaria, Leishmaniasis, sleeping sickness and Chagas's disease. These are all tropical or sub-tropical diseases, listed by WHO among the six major diseases that can be fatal and cause a large number of deaths a year in Asian and African countries. The chromosomes of these pathogens show size polymorphisms among strains and also among homologous chromosomes, and the telomeric regions have highly variable number of repeat sequences. The telomeric region of *P. falciparum* 3D7 has been shown to contain copies of highly virulent genes like rif, var and pfmc-2tm. The plasticity and dynamic nature of the telomeres and subtelomeres allow genes located in these regions to evolve rapidly and adapt to their environment. In this paper we use the versatile technique, Mutual Information from Information Theory, to study the telomeric and subtelomeric sequences from the genomes of above-mentioned pathogens and compare them to telomeric sequences of well known small eukaryotic chromosomes like Sachharomyces cerevisiae, Encephalitozoon cuniculi, Schizosachharoyces pombe as well as linear prokaryotic chromosomes of Borrelia burgdorferi and Streptomyces coelicolor. Mutual Information function, which computes the correlations between nucleotides on the DNA sequence at any given base separation, may be used to analyze the structure of chromosomes. It has been shown elsewhere that occurrence of tandem repeats in the DNA sequence lead to high values of correlation between the repeated nucleotides and hence high value of Mutual Information function. It is shown that the high value of Mutual Information for very large base separations (0.2–0.5 Mb in different cases), is caused by extended repeats of G-rich sequences in the telomeres of eukaryotes, which have been earlier shown to be a requirement for building G-quadraplex structure of some telomeres. Identification of signature sequences of telomeric and sub-telomeric regions of chromosomes of pathogens may lead to development of probable epitopes for drug designing.

061: Disease gene mapping in populations with mixed ancestry

[1]**Arti Tandon**, [3]Nick Patterson, [2]Alkes Price, [4]Simon Myers, [2,3]David Reich

[1]Independent Consultant, Mumbai, India, [2]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston,MA 02115, United States of America, [3]Broad Institute of Harvard and

MIT, 7 Cambridge Center, Cambridge, MA 02142, United States of America, [4]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, OX1 3PS, United Kingdom

We have been focusing in the past few years on developing statistical methods and software which implement them, to map disease genes in samples of mixed ancestry, such as African Americans. Additionally, the group has access to a large number of experimental data sets which have been useful in validating these methods, and in discovering new genetic variants that confer risk for a number of diseases. The initial statistical software developed by the group was ANCESTRYMAP in 2004. This method uses a Hidden Markov Model nested within a Markov Chain Monte Carlo, to infer the ancestry of genomic blocks in admixed samples. Then we use these ancestry estimates to look for association to disease, by focusing on blocks which have unusual ancestry in one of the parental populations compared to the genome wide average. We have successfully implemented this method in African Americans to find genetic variants that increase risk for Prostate Cancer, Multiple Sclerosis, end stage renal disease, and White Blood Cell count. This method has proven to be an extremely useful tool in these cases, and we are now trying to extend its' use to other admixed populations such as Latinos.

However, ANCESTRYMAP requires unlinked markers for analysis, and does not work that well in case of ancient admixture. With the advent of whole genome scans in the past few months, one would like to use all that information to make more accurate local ancestry estimates. These estimates are useful not only in disease mapping, but in understanding human history as well. Currently we are developing a new statistical method HAPMIX. This is a haplotype based method for inferring local ancestry which is based on an explicit population genetic model, and can be used effectively with whole genome scan data. The goal is to have robust methods which can be used to estimate ancestry in challenging scenarios, such as ancient admixture, and when we do not have modern surrogates for the true ancestral populations. We are currently trying to use the various methods in combination to maximize their strengths to look for disease association in admixed populations, and for other applications of local ancestry estimate.

062: PepCluster: a web-tool to annotate bacterial proteins by consensus and core invariant peptide signatures

Bhupesh Taneja, Suchir Arora, Aparna Basu, S. K. Brahmachari

Institute of Genomics and Integrative Biology, Mall Road, Delhi-110007, India

With advancements in automated DNA sequencing strategies over the last few years, there is a vast pool of genomics data available today. It is imperative to exploit this enormous genomics database towards identification and characterization of gene function. This would not only help in understanding their cellular roles but also enable the identification of genes and gene products with potential new and novel functions. Although a number of tools for prediction of protein function based on different levels sequence conservation and motif identification, viz. Pfam, PRINTS, Prosite among many others are available, a large subset of genes and proteins being sequenced remain unannotated. In order to overcome this problem, we have earlier developed an alternative tool for protein identification and annotation called CoPS (Comprehensive Peptide Signature Database) based on invariant sequences present in 52 bacterial genomes (Prakash et al. in Bioinformatics, 2004; Prakash et al. in J Mol Biol, 2005). The 12076 signatures of CoPS have been analyzed for similarities and overlaps and grouped together into functionally homogenous clusters with a corresponding consensus signature for each cluster into a new database, PepCluster. PepCluster rids the redundancy present in some CoPS signatures and contains a reduced set of 5620 conserved signatures. These clusters enable comparison of relationships between the invariant sequences of proteins of similar function present in each cluster. PepCluster now offers a highly useful tool for assigning function to unannotated proteins or alternately also provides functionally important residues present in a protein sequence that may be searched in a user-friendly manner. While assigning function to the unknown protein, the webserver provides detailed information on the sequence of signature peptide, its position and the functional category to which it belongs. Towards user advantage, two functional classification schemes have been embedded to the backend, namely COG's, which was assigned with manual intervention and ARC, which assigns the function obtained for a cluster signature automatically. Details on this webserver and its utilities will be presented.

063: A resource of molecular alterations in breast cancer

[1]**Deepthi Telikicherla**, [1]Kumaran Kandasamy, [1]Renu Goel, [1]Mukhtar Ahmed, [1]Suresh Mathivanan, [1]Devi S Somanathan, [1]Yashwanth Subbannayya, [1]Lakshmi Dhevi N. Selvan, [1]Prathibha Ranganathan, [1,2]Akhilesh Pandey

[1]Institute of Bioinformatics, Bangalore 560066, India, [2]The Johns Hopkins University School of Medicine, Baltimore, MD 21205, India

Cancer pathogenesis usually involves a multitude of molecular alterations at the genomic, transcriptomic and proteomic levels. Recent advances in molecular genetic techniques have led to cataloguing of a large number of molecular alterations in premalignant and malignant forms of breast neoplasms. Most publications focus on individual alterations found in particular types of cancers while a few others are high-throughput studies that report molecular alterations on a global scale. We have developed Breast Cancer Database (BCD) as a resource that includes data from both of these approaches. BCD is a unique repository of known molecular alterations at the DNA, mRNA and protein levels in addition to drug-induced molecular alterations reported in breast cancer. DNA amplification, translocation, deletion, insertion, single nucleotide change and methylation events constitute the genomic alterations. The transcriptomic alterations include mRNA expression level changes and alternatively splice variants in breast cancer. Alterations in the proteome encompass changes in amino acid sequence, altered subcellular localization, altered expression levels, changes in the pattern of post-translational modifications, protein isoforms and alterations in enzymatic activities of proteins. A catalog of drug induced alterations at the DNA, mRNA and protein level is also included in the database. Overall, this resource includes molecular alterations in 2,958 genes found during the course of breast cancer development with a total of 10,267 alterations that have been annotated at the DNA, mRNA and protein levels. BCD is a unique online gene-centric resource that will provide the researcher with a tool to navigate molecular alterations that could be valuable as potential diagnostic, prognostic or therapeutic markers in breast cancer. The Breast Cancer Database is available as a worldwide web resource from http://www.breastcancerdatabase.org/.

064: Dissecting the transcriptional regulation of metastasis

[1]**Ram Krishna Thakur**, [2]Vinod Kumar Yadav, [1]Akinchan Kumar, [1]Richa Basundra, [1]A. Vijay Subbarao, [1]Anirban Kar, [1,2]Shantanu Chowdhury

[1]Proteomics and Structural Biology Unit, Institute of Genomics and Integrative Biology, Mall Road, Near Jubilee Hall, Delhi, India, [2]G.N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Mall Road, Near Jubilee Hall, Delhi, India

The biology of metastasis regulation is poorly understood. Apart from micro-environmental pressures, several transcription factors have been implicated in control of metastasis. Non-metastatic23 (NM23) family of proteins has been associated with suppression of metastasis; however, their association is mechanistically unresolved. Using chromatin immunoprecipitation coupled to human promoter microarrays (ChIP-chip) we performed genome wide location analysis (GWLA) for NM23 H2; a transcription factor involved in the regulation of the proto-oncogene c-MYC. The physical mapping of NM23 H2 across 25,500 human gene promoters indicates wide spread regulatory roles of NM23 H2. Applying de novo motif discovery algorithms, we identified a signature DNA sequence bound by NM23 H2 in cellular conditions. However, there is a complex relationship between transcription factor binding to the DNA motif in promoter and concomitant change in transcription of target genes. Comprehensive bioinformatic analyses associate several other transcription factors which co-operate with NM23 H2 in its transcriptional activity thereby suggesting a regulatory module for metastasis regulation. A profiling of global gene expression changes using oligonucleotide microarrays following siRNA based targeted depletion of NM23 H2 in lung adenocarcinoma cells provides further insights into metastasis regulation. The differentially expressed genes control various biological pathways related to apoptosis, cell cycle progression, cell differentiation, tumor invasion, cell localization and adhesion. Taken together, these whole genome mapping and expression studies together with computational analyses provide novel insights into the complexity of transcriptional regulation of metastasis.

065: Markov segmentation of human X chromosome

[1]**Vivek Thakur**, [2]Ashwin Kelkar, [2]Deepti Deobagkar, [1,3]Ram Ramaswamy

[1]Centre for Computational Biology and Bioinformatics, School of Information Technology, Jawaharlal Nehru University, New Delhi 110 067, India, [2]Department of Zoology, Pune University, Pune 411 008, India, [3]School of Physical Sciences, Jawaharlal Nehru University, New Delhi 110 067, India

Markov chain segmentation is a method of identifying compositionally different subsequences in the input sequence. Via this technique it is possible to partition a DNA sequence into fragments that are homogeneous in nucleotide composition. We have applied this technique to the DNA sequence of the human X chromosome to analyse the compositional structure of the X chromosome. The human X chromosome has been known to have acquired DNA in 4 distinct evolutionary events and is thus believed to be composed of 4 distinct evolutionary strata. These strata are reported to be linearly arranged on the X chromosome, in order of their addition to the X chromosome. The oldest stratum (S1) comprising of the almost the entire 'q' arm of the X chromosome while the newest stratum (S5) being the terminal 'p' arm of the X chromosome. In female mammals all copies of the X chromosome in excess of 1 are transcriptionally inactivated. It has been previously shown that the location of a gene on the X chromosome and its ability to undergo inactivation shows a very good correlation. There is no report on the locations of the precise boundaries between the evolutionary strata on the X chromosome as well as the inactivation domains on the X chromosome. Our method of DNA analysis partitions the X chromosome into domains that are homogeneous in their tri-nucleotide composition. This provides an accurate estimation of the location of stratum boundaries with a high resolution map of boundaries between compositionally different regions on the X chromosome. In addition we find evidence for a new boundary within Stratum 3. The novel boundary partitions the DNA in to two regions where the genes show differences in inactivation status. All the genes on the side of the boundary that is towards the 'p' arm, i.e. the newer regions of the X chromosome escape inactivation while all the genes on the other side of this boundary are subject to inactivation. The genes that are subject to inactivation were also found to have a higher selection pressure as compared to the genes on the other side of the boundary that escape inactivation. Markov segmentation of the human X chromosome provides an extremely valuable insight into the evolution of the X chromosome. Segmentation also provides an interesting perspective in the evolution of X chromosomal genes.

066: Joint association analysis of multiple genes in a biological pathway

[1]**Anbupalam Thalamuthu**, [1]Simone Gupta, [1]Garrett Hor Keong Teoh, [2]Kamila Czene, [2]Per Hall, [1]Edison T. Liu, [1]Jianjun Liu

[1]Genome Institute of Singapore, Singapore 138672, Singapore, [2]Karolinska Institute, Stockholm 17177, Sweden

Multiple genes are known to be involved in the etiology of common diseases. Some of the existing methods for the joint association of multiple genetic variants combine the effect of many variants within a single gene or independent variants from multiple genes. Here we propose a joint association testing methodology to study the effect of multiple variants from several genes. We use gene level attribute to combine the information from multiple genes which enable us to test the joint association of several genes. Further we propose methodology to identify a subset of significantly associated genes. The proposed method can be used for testing the joint effect of several genes in a candidate gene study and can easily be extended to identify important genes in specific biological pathways generated from a Genomewide Association (GWA) studies. We evaluate the performance of the proposed methodology using simulated data sets as well as data sets from a candidate gene study. The results show that the joint effect of multiple genes is more powerful compared single gene analysis.

067: Global gene profiling to predict radiation response in Indian women with cervical cancer

[1]**Asha Thomas**, [2]Umesh Mahantshetty, [3]Kedar Deodhar, [2]Reena Engineer, [2]Shyam K. Shrivastava, [2]K. A. Dinshaw, [1]Rita Mulherkar

[1]ACTREC, Tata Memorial Centre, Kharghar, Navi Mumbai, 410210, India, [2]Radiation Oncology Department, Tata Memorial Hospital, Parel, Mumbai, 400012, India, [3]Pathology Department, Tata Memorial Hospital, Parel, Mumbai, 400012, India

Despite being a preventable disease, cervical cancer claims the lives of almost half a million women worldwide each year. India bears one-fifth of the global burden of the disease, with approximately 130,000 new cases a year. Clinically localized cervical cancer can be effectively ablated using surgical or radiation treatments. For advanced cervical cancer radiotherapy with or without chemotherapy remains the major treatment modality followed despite the low response rate. Despite being in the same stage of the disease responses to the treatment differ among the patients. Monitoring gene expression profiles for genome wide changes in gene expression may provide insights into the molecular finger printing of the diseases or even distinguish responders

and non responders to a particular treatment. For the present study we recruited cervical cancer patients who presented to Department of Radiation Oncology, TMH and were in FIGO Stage IIIb. Biopsies of the tumors prior to starting chemo-radiation were collected in liquid N2 after obtaining their informed consent and stored at -80°C. The patients were followed up for more than 5 years and their clinical history recorded. Cryosectioning of the biopsies was done to determine the percentage of tumour in these tissues and only those samples containing 80% or more of tumour cells were selected for the study. The quality, integrity and quantity of the RNA were checked on Agilent Bioanalyzer. Only samples having good quality RNA were further used for microarray. To identify a set of genes related to radio sensitivity and to establish a predictive method, expression profiles of radiosensitive and radio resistant tumors were compared on Agilent 44 k oligonucleotide microarray chip. About 500 genes were identified that differentiated responders and non responders. Clustering and Tree view analysis of these genes distinctly separated responders and non responders. Candidate genes from this set are being validated.

068: HGVbaseG2P: a central genetic association study database

[1]**G. A. Thorisson**, [1]R. Free, [1]R. Hastings, [1]O. Lancaster, [2]P. Sarmah, [2]S. K. Brahmachari, [2]D. Dash, [1]A. J. Brookes

[1]Department of Genetics, University of Leicester, Leicester, United Kingdom, [2]Institute of Genomics and Integrative Biology, CSIR, Delhi, India

Many research projects utilize the genetic association study principle, but the results of such studies are sub-optimally reported. They tend to be disseminated through diverse and disconnected databases, journals, and meetings, or quite frequently—particularly for negative findings—not reported at all. Consequently, it is very difficult to compare and contrast the results of different studies. The Human Genome Variation Genotype to Phenotype database (HGVbaseG2P) is therefore being constructed to help promote and integrate the world's genetic association study datasets.

HGVbaseG2P expects to go live mid-2008 at the website http://www.hgvbaseg2p.org/.

HGVbaseG2P will focus on summary level genetic association data (i.e., group rather than individual level information), aiming to provide a new publication medium that combines the best features of a database and a scientific journal. The project will strive to actively collate the latest genetic association datasets from all relevant public data resources, and its underlying database and support software will be made freely available to others so they can likewise create compliant databases for such research investigations. It will also devise tools and support structures enabling individual researchers to conveniently deposit their summary-level results directly into the HGVbaseG2P database. All known SNPs, CNVs, and other human genome variants will be incorporated into HGVbaseG2P as a basal DNA layer. Phenotype and association data will then sit on top of this, and be connected according to the standard Phenotype and Genotype Experiment (PAGE) object model that is soon to be published (see http://www.gen2phen.org/ for a link to this and other related standards). The Phenotype layer represents a key innovation, based upon an elegant and compact, though highly robust and flexible, data model that divides the domain into phenotype feature, method, and value components. Innovative ways to encourage people to submit data, and extensive collaboration with journals and funding bodies on this matter, plus unique project funding concepts, are being explored in order to make this project a long-term successful resource.

Acknowledgements: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No 200754, GlaxoSmithKline, the Institute of Genomics and Integrative Biology, and the University of Leicester.

069: Comprehensive analysis of the splice site regions by comparative genomics

Shashi Rekha Thummala

Deparment of Biochemistry, University of Hyderabad, Hyderabad-500 046, Andhra Pradesh, India

With the discovery of the rapid methods of genome sequencing, the interpretation and understanding of biological sequences is not keeping pace with the enormous amount of the data being generated. The availability of the complete genome sequences of different organisms paves the way to study the various characteristic features of their genome organization. Eukaryotes undergo the process of 'RNA splicing', which involves the splicing of introns from heterogenous RNA (hnRNA or pre-mRNA) to form mature mRNA. This process suggests that the nucleotides at/around the splice sites contain the information that is required for the assembly and binding of the spliceosomal proteins onto the splice sites. In order to obtain this information we have carried out a comparative analysis of the donor and acceptor splice site regions in the genes of five different organisms. We have studied the sub-sequences of size six/ten at the splice site (donor/acceptor) regions of these organisms. This analysis suggests that the distribution of their occurrence is approximately exponential. We have observed that the number of unique sub-sequences at the donor region are less than at the acceptor, suggesting that these sub-sequences are more variable at the acceptor region. Our analysis also suggests that the sub-sequences (at the splice sites) of length $\sim$ 6–8 nucleotides with six bases in intron (including the two central, conserved dinucleotides) and two bases in exon are optimal for the efficient assembly and binding of the spliceosomal complex during the process of splicing. A unique donor sub-sequence that occurs with a high frequency is likely to be associated with an unique acceptor site occurring with high frequency. To discover the pattern of association between the donor and acceptor sites, we have used a scoring model. The score pattern obtained by the alignment of nucleotides at the donor region with the acceptor and vice versa suggest that, a single sub-sequence at the donor region have different degree of similarity with the sub-sequences at the acceptor. This signifies that the donor sub-sequences are more crucial in pairing with their corresponding acceptor sub-sequences during the process of splicing. Further more, the rich variability of the donor and acceptor sites generates greater information, which may be useful in understanding the language of DNA at the splice sites.

070: Human chromosomal DNA digestion with restriction endonucleases in vitro and in silico

Victor Tomilov, Murat Abdurashitov, Valery Chernukhin, Danila Gonchar, Sergey Degtyarev

SibEnzyme Ltd., 2/12, Ak. Timakova str, Novosibirsk, 630117, Russia

In the course of Human Genome Project development, the data on whole human DNA sequences become available. The human DNA primary structures are still being updated, but the majority of known data allow performing various theoretical analyses of the genome sequences.

We have developed a technique of in silico analysis of human genomic DNA cleavage at the wide range of nucleotide sequences

based on earlier proposed method of mammalian genomes digestion (Abdurashitov et al.) Distribution diagrams of calculated DNA fragments have been constructed. The comparison of theoretical results of the DNA cleavage at certain nucleotide sequences, which are the recognition sites of the restriction endonucleases (AluI 5′-AGCT-3′, AsuHPI—5′-GGTGA-3′ and 5′-TCACC-3′, Bpu10I—5′-CCTNAGC-3′ and 5′-GCTNAGG-3′, BstDEI—5′-CTNAG-3′, Bst2UI—5′-CCWGG-3′, BstSCI—5′-CCNGG-3′, BstMAI—5′-GTCTC-3′ and 5′-GAGAC-3′, HinfI—5′-GANTC-3′, BssECI—5′-CCNNGG-3′, FauNDI—5′-CATATG-3′, XbaI—5′-TCTAGA-3′, MroXI—5′-GA-ANN NNTTC-3′, KpnI—5′-GGTACC-3′, Msp20I—5′-TGGCCA-3′, AspA2I—5′-CCTAGG-3′) with the experimental results of in vitro digestion with corresponding restriction endonucleases has been carried out.

A similar study of human Alu- and LINE1-repeats digestion has been performed. The diagrams of DNA fragments distribution for both repeats families have been constructed for digestion at recognition sites of the restriction endonucleases mentioned above. Distribution diagrams of human genomic DNA digestion, which results in formation of low molecular weight DNA fragments, correspond to those for Alu-repeats; whereas the digestion, which results in formation of large molecular weight DNA fragments—are similar to those for LINE-repeats. The repeats families fragments diagrams show the good correspondence with the distribution diagrams of the whole genome DNA cleavage.

All theoretical data have been compared to experimental patterns of human DNA hydrolysis with respective restriction endonucleases and the good correspondence for the most of DNA diagrams has been observed.

Reference

Abdurashitov MA, Tomilov VN, Chernukhin VA, Gonchar DA, Kh Degtyarev S Cleavage of mammalian chromosomal DNA by restriction enzymes in silico (Online version—http://science.sibenzyme.com/article14_article_27_1.phtml).

071: Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles

[1]**Ruchi Verma**, [2]Ajit Tiwari, [2]Sukhwinder Kaur, [2]Grish C. Varshney, [1]Gajendra P. S. Raghava

[1]Bioinformatics Centre, Institute of Microbial Technology, Institute of Microbial Technology, Sector 39-A, Chandigarh, India, [2]Cell biology and Immunology, Institute of Microbial Technology, Institute of Microbial Technology, Sector 39-A, Chandigarh, India

Background: Malaria parasite secretes various proteins in infected RBC for its growth and survival. Thus identification of these secretory proteins is important for developing vaccine/drug against malaria. The existing motif-based methods have got limited success due to lack of universal motif in all secretory proteins of malaria parasite. Results: In this study a systematic attempt has been made to develop a general method for predicting secretory proteins of malaria parasite. All models were trained and tested on a non-redundant dataset of 252 secretory and 252 non-secretory proteins. We developed SVM models and achieved maximum MCC 0.72 with 85.65% accuracy and MCC 0.74 with 86.45% accuracy using amino acid and dipeptide composition, respectively. SVM models were developed using split-amino acid and split-dipeptide composition and achieved maximum MCC 0.74 with 86.40% accuracy and MCC 0.77 with accuracy 88.22%, respectively. In this study, for the first time PSSM profiles obtained from PSI-BLAST, have been used for predicting secretory proteins. We achieved maximum MCC 0.86 with 92.66% accuracy using PSSM based SVM model. All models developed in this study were evaluated using fivefold cross-validation technique. Conclusion: This study demonstrates that secretory proteins have different residue composition than non-secretory proteins. Thus, it is possible to predict secretory proteins from its residue composition-using machine learning technique. The multiple sequence alignment provides more information than sequence itself. Thus performance of method based on PSSM profile is more accurate than method based on sequence composition. A web server PSEApred has been developed for predicting secretory proteins of malaria parasites, the URL can be found in the Availability and requirements section.

072: Gene expression analysis studies for acute myeloid leukemia to find out putative drug targets using in-silico approach

[1]**Gulshan Wadhwa**, [2]Rajendra Singh, [3]Priyanka Narad

[1]Department Of Biotechnolgy, Biotechnology Information Centre, Apex Biotechnology Centre, C.G.O. Complex, New Delhi, India, [2]Indian Institute of Veterinary Sciences, IVRI, Izatnagar, India, [3]Birla Institute of Technolgy, BIT Mesra, Ranchi, India

Many cases of hereditary blood cancer are due to mutations in genes or the fusion of BCR-ABL genes. Acute myelogenous leukemia (AMLs) are genetically heterogeneous and characterized by chromosomal rearrangements that produce fusion proteins with aberrant transcriptional regulatory activities. The histopathological changes in these cancers are often characteristic of the mutant gene. We hypothesized that the genes expressed by different types of tumors are also distinctive, perhaps allowing us to identify cases of hereditary blood cancer on the basis of gene expression profiles. RNA from samples of primary tumors from eight carriers of the AMLfail mutation, seven carriers of AMLsuccess were taken for the micro array experiment studies. Statistical analyses were used to identify a set of genes that could were differentially expressed to find out the putative targets of the disease. Permutation analysis of multivariate classification functions established that the gene-expression profiles of tumors with AMLsuccess and AMLfail groups reveal significant differences between the two and thereby we are able to adjudge the differentially expressed genes amongst the data given. The work undertaken was to find out the differential expression of the genes using the tool caGEDA and assess and study each of the output generated. The outputs were seen in the form of graphs and excel data files which could be used to infer and provide statistical details about the data into consideration. The tool is special software to study the expression of cancer genes and it would result in a better understanding of the disease. The main basis of our work was to perform the analysis of microarray data and to infer the various useful results from that data. It has been a major bottleneck in the past to analyze the data and an attempt has been made to perform the analysis to the best of its knowledge.

073: In silico analysis of cloned seed storage protein promoters of wheat, oat and rice for genetic manipulation of triticin gene, a possible target for wheat nutritional quality improvement

[1]**Dinesh Yadav**, [1]Vinay Kumar Singh, [2]Nagendra K Singh

[1]Department of Molecular Biology and Genetic Engineering, College of Basic Sciences and Humanities, G.B. Pant University of

Agriculture and Technology, Pantnagar, 263 145 Uttarakhand, India, [2]National Research Centre on Plant Biotechnology, IARI, Pusa Road, New Delhi-110012, India

Nutritional security is one of the important issues for scientific research worldwide. Several attempts have been made to enhance the nutritional quality of staple crops representing cereals and legumes using the modern biotechnological approach involving transgenic, genomics and proteomics. The identification of genes rich in essential amino acids is still a bottleneck for developing transgenic crops though in vitro regeneration protocols and methods of gene transfer has been optimized in major crops consumed by the majority of the populations in world. Seed Storage protein genes have already been identified as targets for cereals and legumes nutritional quality improvement. Triticin, a minor seed storage protein of wheat accounting for only 5% of total seed storage proteins has been identified to be rich in lysine, a limiting amino acid in wheat crops. Attempts have been made to manipulate the triticin gene by cloning full length cDNA, investigating the inherent variation in the hypervariable region of triticin among different wheat progenitors and cloning seed storage protein promoters from Indian varieties of wheat, oat and rice for enhancing the expression. A total of 24 promoters have been cloned and submitted sequences with assigned accession numbers EF396165-EF 396188 were in silico analyzed for different seed specific cis-regulatory elements.

074: GenePython: a framework for ab initio genome annotation

Pranav C. Yajnik, Anupam Saraph, Sohan P. Modak

Open Vision, 759/75 Deccan Gymkhana, Pune 411004, India

Current ab initio gene-finding algorithms are based on complex probabilistic models of gene structure. We propose a simpler approach based on the actual set of molecular events that occur as genes are transcribed, spliced and translated. In this approach, sets of signal sequences are identified iteratively to create a pipeline of checks wherein each successive iteration refines the set of putative genes. GenePython is a set of Python classes designed to emulate the behavior of key participants involved in the flow of genetic information. GenePython objects fall into two categories: Sequence objects like DNA, RNA etc. or Virtual Enzymes like RNA Polymerase, Spliceosome, Ribosome etc. The search pipeline requires a 'catalytic action' of Virtual Enzymes on their substrate sequence objects. GenePython objects have a logical structure and an intuitive interface, which, along with the fact that the objects are part of a programmable environment, facilitate easy manipulation of the structure of the pipeline to suit specific needs. GenePython Sequence objects offer an alternative to the commonly used flat-file method of storing sequence data and facilitate storage, retrieval and manipulation of large volumes of sequence data. A GenePython program simulates the flow of genetic information. The rules for the action of the Virtual Enzymes are explicitly coded and completely transparent. For instance, the RNA Polymerase Virtual Enzyme scans a DNA Sequence for signal elements like TATA box, Polyadenylation signals etc. and generates a list of Virtual Transcripts. Similarly, the Ribosome Enzyme would scan a Virtual Transcript for in-frame start and stop codons to translate putative ORFs into Virtual Polypeptides. The output generated by the action of these Enzymes is biologically intuitive and easier to understand than that obtained by probabilistic parsing of DNA sequences. While such explicit coding may lower the sensitivity of the algorithm, GenePython output may prove to be a useful starting point for other gene-finding algorithms by providing them with a highly reduced search set as compared to an entire genome. Python is an easy-to-learn and user-friendly language. GenePython retains this ideology and is a step forward towards making genome annotation an incisive, inclusive and interactive process.