Genomic Med. (2008) 2:177–179
DOI 10.1007/s11568-009-9089-2

ABSTRACTS

Computational biology and structural proteomics

033: High quality manual genome annotation at WTSI

J. E. Loveland, C. A. Steward, J. P. Almeida, I. Barnes, D. R. Carvalho-Silva, A. Frankish, J. G. R. Gilbert, L. Gordon, E. Hart, J. M. Mudge, C. Snow, M. M. Suner, S. Trevanion, L. Wilming, J. L. Harrow, T. Hubbard

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus,
Hinxton, Cambs, UK

The HAVANA group from the Wellcome Trust Sanger Institute is responsible for the manual annotation of coding, transcript and pseudogene loci on human, mouse and zebrafish finished genomic sequence. The total number of protein coding genes and the extent of alternative splicing of loci on the human genome is still unclear. In collaboration with Ensembl, RefSeq at NCBI and UCSC, the CCDS project (Consensus CoDing Sequence) is working to define a core set of protein coding transcripts. Initially limited to human, the project was recently extended to include mouse. Any CCDS candidate transcripts where there is disagreement between the collaborators are manually re inspected, discussed and, where possible, an agreement is reached on a structure. The end result is a combined, non-redundant gene set that is an ongoing activity with the resolution of differences and the refinement of the gene set between CCDS update cycles.

The HAVANA group is also leading the ENCyclopedia Of DNA Elements (ENCODE) scale-up of the gene annotation project (GENCODE). The project includes seven other partner institutes and will integrate expert manual annotation, computational predictions and targeted experimental analysis to generate a complete reference gene set for the whole genome. This will also include the analysis of pseudogenes, experimental validation of putative/novel genes and examination of the protein-coding potential of genes using comparative and structural analysis. This collaboration will also feed into the CCDS project and provide the genomic community with an accurate gene catalogue for the human genome.

All manual annotation produced by HAVANA group is displayed on the VErtebrate Genome Annotation (VEGA) database (http://vega.sanger.ac.uk/).

034: Identification and analysis of novel repeats and domains in human proteome

Lalitha Guruprasad, Rao Satyanarayana, Hema Latha Golaconda

School of Chemistry, University of Hyderabad, GachiBowli Road,
Hyderabad 500046, India

A predominant part of human genome consists of repetitive sequences, encompassing large segmental duplications, interspersed transposon derived repeats and tandem repeats. Amino acid repeats, known, as homopolymeric tracts are present in nearly one-fifth of human gene products. The uncontrolled expansion of trinucleotide repeats in human coding sequences is associated with several neurodegenerative disorders. Examples are Huntington's disease and dentatorubropallidoluysian atrophy, both associated to abnormally long expansions of CAG runs encoding polyglutamine tracts. Repeat structures in humans have been found to play vital roles in various biological functions such as signal transduction, apoptosis, transcription regulation and several diseases.

Realizing the importance of amino acid repeats in proteins, we undertook the study of identifying the novel amino acid sequence repeats and domains in human proteome. We have implemented repeat identification method TRUST and automated the database searching methods that could be applied to a complete proteome. Using these methods for human proteome, we identified seven domains and 18 repeats that have not been reported so far. Repeats in proteins comprise less than 55 amino acid residues and often present in multiple copy numbers and in tandem. All the repeats are required for correct folding and function of protein. Whereas, domains have greater than 55 amino acid residues and present as single or multiple copy numbers in proteins. Domains are structural and functional units and are independent of the rest of the protein. We have considered only those repeats with length greater than 25 amino acid residues in this work. Our analysis suggests that some of the repeats and domains identified in this work are associated with diseases such as polycystic kidney disease and Williams-Beuren syndrome. In this presentation, we discuss our automated methods for the identification of novel repeats and domains from any proteome. We further report our findings of novel repeats and domains from the human proteome analysis and suggest their functional importance and role in human diseases.

035: Defining expression signatures of known cancer genes using seriation analysis of SAGE libraries from Cancer Genome Anatomy Project (CGAP)

Olena Morozova, Vyacheslav Morozov, Martin Hirst, Marco Marra

Genome Sciences Centre, British Columbia Cancer Agency,
Suite 100, 570 West 7th Avenue, Vancouver, BC V5Z 4E6, Canada

Cancer is a genomic disease involving many types of molecular aberrations. Recently, a number of cancer genome re-sequencing studies have identified an abundance of sequence variation in cancer samples. Since most of this variation is neutral, a primary challenge of modern cancer genomics has been to distinguish causative cancer mutations from the abundance of neutral polymorphisms. Evolutionary models wherein causative sequence variants are presumed to be under positive selection in somatic neoplasms have been adopted for this purpose. However, these models do not take into account the complex scope of a cancerous phenotype, including gene expression changes, and thus represent an initial development in the emerging field of cancer genomics. In order to develop more advanced models for predicting causative cancer variants (cancer drivers), a better understanding of the interplay of sequence variation and other phenotypic characteristics of cancer cells is required. To further our understanding of the role of gene expression in defining cancer driver mutations, we examined gene expression patterns of known cancer genes from the Cancer Gene Census (http://www.sanger.ac.uk/genetics/CGP/Census/) in human cancer samples represented in the Cancer Genome Anatomy Project (http://cgap.nci.nih.gov/) resource. The cancer genes in the Census were defined as such based on the clinical implications of mutations in their coding sequence in cancer patients. To identify common expression signatures among cancer genes, we used a novel seriation algorithm that reorders genes based on the similarity of their Serial Analysis of Gene Expression (SAGE) profiles. The seriation of cancer genes expressed in hematopoietic tissues showed that the genes were either consistently up or consistently down regulated in acute myelogenous leukemia cancer samples versus normal samples. In contrast, seriation analysis of Illumina sequence tag libraries from melanoma and normal skin revealed inconsistencies in the expression patterns of cancer genes in melanoma samples. The finding of characteristic expression signatures of cancer genes in a number of cancer samples of the same tumor type implies that directional selection of gene expression, in addition to that of DNA sequence, occurs in some cancer systems. Therefore, we suggest that gene expression information should be incorporated into the evolutionary models of cancer that currently use sequence data alone.

036: Structure–function correlations in LuxS from bacteria: analysis of protein–protein interface clusters by graph theoritical approach

Moitrayee Bhattacharyya, Saraswathi Vishveshwara

Indian Institute of Science, Molecular Biophysics Unit, Bangalore, India

The genome of a wide variety of prokaryotes contains the luxs gene homologue, which encodes for the protein S-ribosylhomocysteine lyase (LuxS). This protein is responsible for the production of the quorum sensing molecule, AI-2, which coordinates changes in behaviour as a function of cell density. Very often quorum sensing has been directly associated to pathogenicity. But in some prokaryotes no pathogenic role has been attributed to quorum sensing and it has been found to have an effect on metabolism, and thus the overall fitness and 'well being' of the organism. In some organisms, the LuxS has a role in controlling flagellar morphogenesis and it was also known to modulate the toxin production in some virulent bacteria. So it is a single protein performing diverse roles. In our study, we have considered LuxS from a variety of bacterial species. Some of the protein structures were modelled due to the lack of crystallographically available structures. The interface which contains the active site was known to be structurally and functionally important. Thus, the protein structure network (PSN) graphs were constructed to characterize the interface of this homodimeric protein. Our analysis revealed potential sites of mutation and geometric patterns on the interface that was not evident from conventional sequence alignment studies. The key features presented by the protein interface is being investigated for the classification of the proteins in relation to their function. This sort of characterization enables a better understanding of the relation between the luxs gene and its functional role in the prokaryotes.

037: Intrinsic versus induced effects in DNA structure

[1]**Manju Bansal**, [1]Deepti Karandur, [2]Nikhila Jayaprakash, [1]Arvind Marathe

[1]Molecular Biophysics Unit, Indian Institute of Science,
Bangalore-560012, India, [2]Department of Biotechnology,
Indian Institute of Technology, Chennai-600036, India

An important question of biological relevance is the polymorphism of the double-helical DNA structure in its free form, and the changes that it undergoes upon protein-binding. We have analysed a database of free DNA crystal structures to assess the inherent variability of the free DNA structure and have compared it with a database of protein-bound DNA crystal structures to ascertain the protein-induced variations. Most of the dinucleotide steps in free DNA display high flexibility, while protein binding prefers the duplex in B-DNA conformation and in certain cases, causes the DNA backbone to attain energetically unfavourable conformations. At the gross structural level, several protein-bound DNA duplexes assume a curved conformation in the absence of any large protein-induced distortions, indicating that a series of normal structural parameters at the dinucleotide and trinucleotide level, similar to the ones in free B-DNA, can give rise to curvature at the overall level. Thus the free DNA molecule, even in the crystalline state, samples a large amount of conformational space without the aid of any large ligands. For most of the bound DNA structures, across a wide variety of protein families, the average parameters are quite close to the free 'B-like' DNA oligomer values, indicating that in a large number of cases, protein binding may be the result of the protein 'finding' a suitable stretch of genomic DNA in the 'right' conformation, at that time point and then locking it in that conformation for a certain time period.

The crystal structure database analysis was complemented by molecular dynamics studies on the quorum sensing transcription factor, traR, bound to the trabox DNA sequence-d ($A_1T_2G_3T_4G_5C_6A_7G_8A_9T_{10}C_{11}T_{12}G_{13}C_{14}A_{15}C_{16}A_{17}T_{18}$). Simulations have also been carried out on an unbound trabox and a mutated trabox sequence. The central spacer region d (A_7–T_{12}) in the unbound, unmutated trabox is relatively straight with a characteristic narrow minor groove, a conformation similar to the one observed in the crystal structure complex. The unbound, unmutated trabox samples a wide range of bent conformations, including the one observed in the traR–trabox complex crystal structure. Mutations G_8>C and C_{11}>G introduce kinks around the CA/TG steps that distort the spacer region, and affect the groove width and the overall conformation of the trabox, possibly making it unfavourable for binding by either or both the monomers of the traR. Further analysis is being carried out.

038: Three-dimensional structure of *Vibrio cholerae* hemolysin oligomer by cryoelectron microscopy

Somnath Dutta, Budhaditya Mazumdar, K. K. Banerjee,
A. N. Ghosh

National Institute of Cholera and Enteric Diseases, P -33, C.I.T. Road,
Scheme-XM, Beleghata, Kolkata-700010, India

Vibrio cholerae hemolysin (HlyA) is an extra cellular membrane damaging protein. Molecular weight of hemolysin (HlyA) is 65000 Da. This protein exists in two stable states, a water-soluble monomer and an oligomeric integral membrane protein. The protein is synthesized as an 82 kDa preprohemolysin by *V. cholerae* EI Tor 01 and non-Ol strains and exported to the culture medium as the 79 kDa prohemolysin (proHlyA). Proteolytic removal of the 132-residue N-terminal stretch generates the mature 65 kDa HlyA with a specific hemolytic activity. HlyA transforms itself in contact with target biomembranes and synthetic lipid vesicles containing cholesterol into water filled transmembrane heptameric channels of internal diameter 1.5 nm. Recent transmission electron micrographic study has revealed that HlyA binds to cholesterol and forms oligomers at the interface of cholesterol and water. The aim of the present study is to determine the three-dimensional structure of 65 kDa hemolysin using cryoelectron microscopy and single particle methods. Holey carbon grid was glow-discharged and 65 kDa hemolysin sample was placed on the surface of the holey grid. After blotting nearly to dry with a piece of filter paper, the grid was plunged into liquid ethane at −180°C and protein molecules were embedded in vitreous ice. The grid was placed on a GATAN 626DH cryo-holder and inserted into FEI Tecnai 12 BioTwin transmission electron microscope. Electron micrographs of frozen hydrated HlyA were taken at 120 kV. Images were taken at different defocus values using 'low dose' software. These micrographs were digitized using Nikon Coolscan 9000ED film scanner. The three-dimensional structure of Hly A was determined using the EMAN 1.7 software operating on Linux Fedora Core 4 platform. 3200 HlyA particles were selected using 'boxer' program of EMAN software. Contrast Transfer Function (CTF) correction was performed using 'ctfit' program of EMAN software. Rotational symmetry was assessed initially in EMAN using 'startcsym'. Finally three-dimensional reconstruction of 65 kDa *V. cholerae* hemolysin was performed using 'refine' program of EMAN software. The resolution was determined by using a 0.5 Fourier Shell Coefficient (FSC) threshold. The 0.5 value of Fourier Shell Coefficient (FSC) indicates a final resolution was bellow 25 Å. Three-dimensional image of 65 kDa *V. Cholerae* hemolysin was visualized using CHIMERA software.