SNP special interest group

SNP-SIG Meeting

Identification and annotation of SNPs in the context of structure, function, and disease.

ISMB/ECCB 2011
July 15th 2011, Vienna (Austria)

http://snps.uib.es/snp-sig

Highlight Speakers

Atul J. Butte
Stanford University, Stanford (CA), USA

Clinical Assessment Incorporating a Personal Genome.

Mauno Vihinen,
Tampere University, Tampere, Finland.

Genetic variations: origin, effects and prediction.

Keynote Speakers

Steven Brenner
University of California, Berkeley (CA), USA

CAGI Experiments.

Burkhard Rost
Technische Universitat, Munchen, Germany

Trivial step from predicting the effects of SNPs to medicine.

SNP-SIG Organizers

Yana Bromberg, Rutgers University, New Brunswick (NJ), USA
Emidio Capriotti, Stanford University, Stanford (CA), USA

Poster Session

Janita Thusberg, Buck Institute, Novato (CA), USA

Roundtable Discussion

Chris Baker, University of New Brunswick, Saint John (NB), Canada
Maricel Kann, University of Maryland, Baltimore (MD), USA
Sean Mooney, Buck Institute, Novato (CA), USA

SNP-SIG Meeting Programme - July 15th 2011, Vienna (Austria)

08:30 – 08:45 Welcome from the committee

SI: Annotation & prediction of structural/functional impacts of coding SNPs

08:45 – 09:40 **Highlight Speaker: Mauno Vihinen**, Tampere University (Finland)
Genetic variations: origin, effects and prediction.

09:40 – 10:05 **Christian Schaefer**. Technische Universitat Munchen (Germany)
Can we predict structural change upon point mutation?

10:05 – 10:30 **Gilad Wainreb**. Tel Aviv University (Israel)
Protein stability: A single recorded mutation aids in predicting the effects of other mutations in the same amino acid site.

10:30 – 10:45 Coffee Break

10:45 – 11:10 **Piero Fariselli**. University of Bologna (Italy)
Predicting cancer-associated germline variations in proteins.

11:10 – 11:35 **Alain Laederach**. University of North Carolina, Chapel Hill (USA)
Effects of disease-associated SNPs on the structure of the transcriptome.

11:35 – 12:10 **Keynote: Burkhard Rost**, Technische Universitat Munchen (Germany)
Trivial step from predicting the effects of SNPs to medicine.

12:10 – 12:25 **Company Presentation: Frank Schacherer,** BIOBASE GmbH.
Manually curated databases for SNP analysis.

12:25 – 13:30 **Lunch Break and Poster Session with the Authors**

SII: SNPs & Personal Genomics: GWAS, populations and phylogenetic analysis

13:30 – 14:25 **Highlight Speaker: Atul J. Butte,** Stanford University (USA)
Clinical Assessment Incorporating a Personal Genome.

14:25 – 14:50 **Konrad Karczewski**. Stanford University (USA).
Assessing Functional and Clinical Significance of Regulatory Variants.

14:50 – 15:15 **Joel Dudley**. Stanford University (USA).
Evolutionary meta-analysis reveals ancient constraints affecting missing heritability and reproducibility in disease association studies.

15:15 – 15:30 Coffee Break

15:30 – 15:55 **Adam Frankish**. Wellcome Trust Sanger Institute, Hinxton (UK)
Is Understanding Alternative Splicing Important in Interpreting the Potential Functional Effects of SNPs?

15:55 – 16:20 **Sudhir Kumar**. Arizona State University, Tempe (USA).
Comparative Genomics as an Evolutionary Telescope for Genomic Medicine to Peer into the Universe of Human Mutations.

16:20 – 16:55 **Keynote: Steven Brenner,** University of California, Berkeley (USA).
CAGI Experiments

16:55 – 17:45 **Round Table Discussion**

17:45 – 18:00 Closing remarks from the committee

Highlight Presentations

SNP-SIG Meeting – ISMB/ECCB 2011, July 15th Vienna (Austria)

GENETIC VARIATIONS: ORIGIN, EFFECTS AND PREDICTION

Mauno Vihinen*

Tampere University, Finland

email: mauno.vihinen@uta.fi

Each individual has millions of genetic variations in their genome. This information can now be efficiently revealed with next generation sequencing techniques. As some variations are associated with diseases it would be important to be able to identify them. Due to the large amount of data this is not possible with experimental methods. Computational approaches are needed to study and analyze variations and to prioritize likely pathogenic or neutral cases. The origin of genetic variations and mechanisms behind them will be discussed. These concepts are needed to be able to systematically classify variations and their effects. Variation Ontology (VariO) is developed to facilitate systematic variation effect description. Many existing methods are devoted to recognizing harmful variations. This is an important prediction task, which can be extended to study of consequences and effects of harmful variants. Performance of certain prediction tools will be discussed based on large scale tests of experimentally verified cases. Our experiences from the use and development of prediction tools as well as variation database curation and development will be discussed with view for the future.

CLINICAL ASSESSMENT INCORPORATING A PERSONAL GENOME

Atul J Butte*

* Stanford University, USA*

email: abutte@stanford.edu

Dr. Butte's lab at Stanford builds and applies tools that convert more than 30 billion points of molecular, clinical, and epidemiological data measured by researchers and clinicians over the past decade into insights into diagnostic and therapeutic potential. Dr. Butte, a bioinformatician and pediatric endocrinologist, will highlight his recent work on the first clinical evaluation of a patient presenting with a personal genome.

Selected Presentations

SNP-SIG Meeting – ISMB/ECCB 2011, July 15th Vienna (Austria)

GENETIC VARIATIONS: ORIGIN, EFFECTS AND PREDICTION

Christian Schaefer*, Burkhard Rost*

*Technical University Munich. Germany

email: schaefer@rostlab.org, rost@in.tum.de

Non-synonymous single nucleotide polymorphisms (nsSNPs) refer to point mutations in coding regions of the DNA that change the amino acid within the gene product. Once the protein gets expressed, nsSNPs could have an effect on its stability and/or function or not. There exists a plethora of methods that predict the outcome of an occurring nsSNP. Usually, change in protein stability is deduced from the change of free energy (ddG) between wildtype and mutant structure imposed by the amino acid exchange. Here we propose a more direct way to measure structural change. We compile a data collection out of the Protein Data Bank (PDB) consisting of a set of structurally superimposed protein fragment pairs that share identical flanking regions in sequence and one mismatch amino acid in their center. By measuring the root mean square displacement (RMSD) between two fragments, we infer the structural effect posed by the central mismatch. With the dataset at hand, we aim to machine-learn the structural displacement due to a nsSNP by sequence-derived features alone. We tackle that problem as follows: By setting discrete RMSD thresholds to define structural effect/no-effect, we view the given task as a classification problem.

PROTEIN STABILITY: A SINGLE RECORDED MUTATION AIDS IN PREDICTING THE EFFECTS OF OTHER MUTATIONS IN THE SAME AMINO ACID SITE

Gilad Wainreb*, Lior Wolf , Haim Ashkenazy, Yves Dehouck, Nir Ben-Tal*

* Tel Aviv University, Israel

email: wainreb@tau.ac.il, nirb@tauex.tau.ac.il

We introduce an approach for predicting the change in a protein's stability that arises from a single-site amino acid substitution starting from sequence or structure . The approach uses available data on mutations occurring in the same position and in other positions. Our algorithm, named Pro-Maya (Protein Mutant stAbilitY Analyzer), combines a newly developed collaborative-filtering–based algorithm, Random Forests regression, and a diverse set of descriptors. The algorithm was benchmarked on two previously utilized datasets of mutations and on a (third) validation set. The results indicate that using known $\Delta\Delta G$ values of mutations at the query position improves the accuracy of $\Delta\Delta G$ predictions for other mutations in that position. Our predictions in such cases are significantly more accurate than those of similar methods, achieving, e.g., a Pearson correlation coefficient of 0.79 on the validation set.

Pro-Maya is freely available via web server at http://bental.tau.ac.il/ProMaya.

PREDICTING CANCER-ASSOCIATED GERMLINE VARIATIONS IN PROTEINS

Piero Fariselli* , Eva Balzani, Pier Luigi Martelli, Rita Casadio*

*University of Bologna. Italy
email: piero@biocomp.unibo.it, casadio@biocomp.unibo.it

Various computational methods are presently available in order to classify whether a protein variation is disease-associated or not. However data derived from recent technological advancements make it feasible to extend the annotation of disease-associated variations in order to include specific phenotypes. By this, a better characterization of the genes that may increase the risk of inherited Mendelian diseases is possible. Here we describe a new implementation based on support vector machines that takes as input the protein variant and the protein function, as described by its associated GO terms. When a cross-validation approach is adopted, the method well discriminates within germline variants those that are likely to be cancer-associated, scoring with 90% accuracy and 0.61 Matthews correlation coefficient on a set comprising 6478 germline variations (16% are cancer-associated) in 592 proteins. Furthermore the method is capable of correctly excluding some 96% of 3392 somatic cancer-associated variations in 1983 proteins not included in the training/testing set.

EFFECTS OF DISEASE-ASSOCIATED SNPS ON THE STRUCTURE OF THE TRANSCRIPTOME

Matthew Halvorsen, Joshua Martin, Lauren Neulander, Wes Sanders, Art Beauregard, Justin Ritz, Alain Laederach*

* University of North Carolina, USA
email: alain@unc.edu

Given the high percentage of the genome that is transcribed, we postulate that for some observed genetic associations the disease phenotype is caused by a structural rearrangement in a regulatory region of the RNA transcript. To identify such mutations we have performed a genome wide analysis of all known disease-associated Single Nucleotide Polymorphisms (SNPs) from the Human Gene Mutation Database (HGMD) that map to the untranslated regions (UTRs) of a gene. Rather than using minimum free energy approaches (e.g. mFold), we use a partition function calculation that takes into consideration the ensemble of possible RNA conformations for a given sequence. We identified in the human genome over 50 disease-associated SNPs that significantly alter the global conformation of the UTR to which they map. For six disease-states (Hyperferritinaemia Cataract Syndrome, β-Thalassemia, Cartilage-Hair Hypoplasia, Retinoblastoma, Chronic Obstructive Pulmonary Disease (COPD), and Hypertension) we identified multiple SNPs in UTRs that alter the mRNA structural ensemble. Using a Boltzmann sampling procedure for sub-optimal RNA structures, we are able to characterize and visualize the nature of the conformational changes induced by the disease-associated mutations in the structural ensemble. We observe in several cases (specifically the 5' UTRs of FTL and RB1) SNP induced conformational changes analogous to those observed in bacterial regulatory Riboswitches when specific ligands bind. We propose that the UTR and SNP combinations we identify constitute a "RiboSNitch," that is a regulatory RNA in which a specific SNP has a structural consequence that results in a disease phenotype. Using our Capillary Automated Footprinting Analysis (CAFA) approach, we performed high-throughput SHAPE chemical mapping on the FTL 5' UTR (wild-type and 6 disease-associated mutations) to confirm the existence of a human RiboSNitch in the FTL gene.

ASSESSING FUNCTIONAL AND CLINICAL SIGNIFICANCE OF REGULATORY VARIANTS

Konrad J. Karczewski*, Joel T. Dudley, Nicholas P. Tatonetti, Russ B. Altman, Atul Butte, Michael Snyder*

*Stanford University, USA
email: konradjk@stanford.edu, mpsnyder@stanford.edu

Many genomic variants are discovered outside of genes, where their functional consequences are more difficult to characterize. However, as many of these variants are associated with disease1, it is likely that they affect molecular physiology at the level of gene regulation. We investigate the role of variants in regulatory regions, on both transcription factor cooperativity as well as disease pathophysiology. First, we developed the Allele Binding Cooperativity (ABC) test and the ALPHABIT pipeline, which utilizes variation in transcription factor binding among individuals to discover combinations of factors and their targets. We find some factors that have been known to work with NF!B (E2A, STAT1, IRF2), but whose global co-association and sites of cooperative action were not known, and discover one co-association (EBF1) that had not been reported previously. Second, we demonstrate a systematic approach to combine disease association, transcription factor binding, and gene expression data to assess the functional consequences of variants associated with hundreds of human diseases. We find that disease-associated SNPs are enriched in NF!B binding regions overall, and specifically for inflammatory mediated diseases, such as asthma and atherosclerosis. Using genome-wide binding variation information, we find regions of NF!B binding correlated with disease-associated variants in an allele-specific manner. Furthermore, we show that this binding variation is often correlated with expression of nearby genes, which are also found to have altered expression in independent profiling of the variant-associated disease condition. In this systematic approach, we close a major loop in biological context-free association studies and assign putative function to many disease-associated SNPs.

EVOLUTIONARY META-ANALYSIS REVEALS ANCIENT CONSTRAINTS AFFECTING DISCOVERY AND REPRODUCIBILITY IN DISEASE ASSOCIATION STUDIES

Joel T Dudley*, Rong Chen, Atul J Butte, Sudhir Kumar

*Stanford University, USA
email: jdudley@stanford.edu

Genome-wide disease association studies contrast genetic variation between disease cohorts and healthy populations to discover single nucleotide polymorphisms (SNPs) and other genetic markers revealing underlying genetic architectures of human diseases. Despite many large efforts over the past decade, these studies are yet to identify many reproducible genetic variants that explain significant proportions of the heritable risk of common human diseases. Here, we report results from a multi-species comparative genomic meta-analysis of 6,720 risk variants for more than 420 disease phenotypes reported in 1,814 studies, which is aimed at investigating the role of evolutionary histories of genomic positions on the discovery, reproducibility, and missing heritability of disease associated SNPs (dSNPs) identified in association studies. We show that dSNPs are disproportionately discovered at conserved genomic loci in both coding and non-coding regions, as the effect size (odds ratio) of dSNPs relates strongly to the evolutionary conservation of their genomic positions. Our findings indicate that association studies are biased towards discovering rare variants, because strongly conserved positions only permit minor alleles with lowest frequencies. Using published data from a large case-control study, we demonstrate that the use of a straightforward multi-species evolutionary prior improves the power of association statistics to discover SNPs with reproducible genetic disease associations. Therefore, long-term evolutionary histories of genomic positions are poised to play a key role in reassessing data from existing disease association studies and in the design and analysis of future studies aimed at revealing the genetic basis of common human diseases.

IS UNDERSTANDING ALTERNATIVE SPLICING IMPORTANT IN INTERPRETING THE POTENTIAL FUNCTIONAL EFFECTS OF SNPS?

Adam Frankish*, Daniel MacArthur, Chris Tyler-Smith, Jennifer Harrow

*Wellcome Trust Sanger Institute, UK
email: af2@sanger.ac.uk

The ultimate goal for studies that identify variation in the human genome is to accurately predict the effects of variants on phenotype. We propose that knowledge of alternative splicing (AS) is essential in providing context for assessment of functional effect and may identify AS transcripts that affect the interpretation of functional impact. For example: an AS transcript skipping an exon can lead to the exclusion of the SNP from the mature mRNA; a novel AS transcript can putatively "rescue" the functional potential of a locus by encoding a transcript with less severe predicted effects; identification of an AS transcript can allow a functional prediction to be made where not previously possible. As such, the comprehensiveness of the variant's impact can be assessed, determining whether its effects should be considered at the level of the individual transcript or the whole locus. When considering the impact of AS it is important that the most accurate and complete geneset available is utilised. The GENCODE consortium are producing the reference gene annotation for the human ENCODE project. Manual GENCODE annotation is generally built using EST, mRNA or protein support; however, RNA_seq data has recently been introduced into the annotation pipeline. As part of the analysis of the Loss of Function variants for the 1000 Genomes project we have annotated and analysed genes containing 885 variants. Of the 597 variants confirmed to affect protein-coding loci, 213 were subject to AS, confirming the importance of AS in functional prediction.

COMPARATIVE GENOMICS AS AN EVOLUTIONARY TELESCOPE FOR GENOMIC MEDICINE TO PEER INTO THE UNIVERSE OF HUMAN MUTATIONS

Sudhir Kumar*

* Arizona State University, USA
email: s.kumar@asu.edu

De novo mutations and single nucleotide variants (SNPs) are being detected with remarkable ease in personal genomes owing to landmark advances in sequencing technologies. With these advances, Genomic Medicine is confronted with answering practical questions about the effect of personal SNPs on individual health. What do these SNPs foretell about my own predisposition to diseases? How do they relate with my current and future health conditions? Will my personal SNPs, and their combinations, affect my response to medical drugs and treatments? However, each individual genome contains millions of mutations, both germ line and somatic, with varying complexities. It is currently impractical to study the effects of such an enormous number of mutations in the experimental laboratory. Even in the 1-3% of the genome that codes for proteins (the Exome), and for which we have some understanding of the function and influence on diseases, an average person carries ~10,000 amino acid SNPs. Fortunately, nature has been experimenting with the effects of new protein mutations over millions to billions of years. In nature's test tube, the process of selection has repeatedly eliminated protein mutants with negative effects on function. Outcomes of these experiments are revealed by comparative genomics of humans with great apes and earliest animals, and beyond. In this discussion, I will discuss how two of these outcomes – long-term evolutionary conservation and multispecies substitutions – constitute two lenses in an evolutionary telescope, which is poised to be a powerful tool in the quest of biomedical scientists to evaluate consequences of the ever expanding universe of human variations.

Poster Session

SNP-SIG Meeting – ISMB/ECCB 2011, July 15th Vienna (Austria)

CLUSTERING DISEASE CONNECTIONS USING DMDM: DOMAIN MAPPING OF DISEASE MUTATIONS

Maricel Kann*, Nathan Nehrt, Thomas Peterson

University of Maryland, Baltimore County. USA

email: mkann@umbc.edu

Domain mapping of disease mutations (DMDM) is a database in which each disease mutation can be displayed by its gene, protein or domain location. By aggregating disease mutations and polymorphisms from all proteins containing a given domain, DMDM's unique domain view highlights molecular relationships between different diseases that might not be observed with traditional gene-centric visualization tools. To examine the relationships between diseases connected at varying levels of specificity (ex. between Parkinson disease and Severe Combined Immunodeficiency, or more generally between neurological and immunological disorders), we are now developing methods for clustering diseases based on their protein domain associations. To benchmark these methods, we performed an initial clustering of disease names. We are currently evaluating methods to cluster diseases by their phenotype similarity in order to group diseases with similar molecular mechanisms and to assess our domain-based cluster methodology. Our website, DMDM, is available at http://bioinf.umbc.edu/dmdm and can be used to identify protein domain sites with high incidence of disease mutations.

COMPREHENSIVE NSSNP DATABASE WITH PREDICTIONS AND ANNOTATIONS OF IMPACT

Christian Schaefer*, Yana Bromberg, Burkhard Rost

Technical University Munich. Germany

email: schaefer@rostlab.org

In the past few years, a plethora of databases emerged containing single nucleotide polymorphisms (SNPs); dbSNP may be the most popular one. Non-synonymous SNPs (nsSNPs) change the amino acid in the gene product and either could have an effect on protein structure and/or protein function or are neutral in that sense. Most nsSNPs, however, lack experimental annotations about their functional impact and most human nsSNPs also have no annotation about their disease-associated role. Here, we introduce a rich database based on dbSNP. While data from human are clearly the most prevalent, we also report nsSNPs from 13 other organisms. Apart from data taken from dbSNP, we also provide nsSNPs from other databases like SwissProt, OMIM and PMD. Amino acid annotations from SwissProt around observed nsSNPs augment the available information. The impact on function of each nsSNP is predicted using SIFT and SNAP. The modular design of database and management scripts allows for easy integration of further databases and predictors, a web front end offers convenient retrieval of information.

FINDING DETERMINANTS OF KINASE SPECIFICITY AND KINASE-SUBSTRATE NETWORK REWIRING IN CANCER

Pau Creixell*, Rune Linding

* *Center for Biological Sequence Analysis, Denmark*
email: paucreixell@gmail.com

Despite the increasing number of cancer genomes sequenced, our interpretation of how genetic mutations perturb signalling networks pushing the cell towards a cancerous phenotype is still limited. Mutations in cancer can perturb and rewire kinase-substrate networks by affecting the so-called determinants of specificity, a largely unidentified group of residues in the kinase domain that regulate substrate specificity. Here, by developing and deploying a novel information-theoretic approach, several of these determinants of specificity present in the kinase domain of all eukaryotic protein kinases are uncovered. The vast majority of determinants of specificity identified by our method were located in the C-lobe, consistent with its role in substrate binding, and often in or around key functional elements such as the catalytic loop or the activation segment, suggesting that kinase domain activation and substrate specificity could be closely related. After integration with mutation data, the identification of these determinants will shed new light on how mutations may facilitate signalling network rewiring by promoting protein kinases to change their specificity and phosphorylate different substrates. Our results could be an important contribution to our effort to understand kinase-substrate networks and how they are rewired in complex diseases.

POSSIBILITIES AND LIMITATIONS FOR THE PREDICTION OF DISEASE RELATED MUTATIONS IN THE HUMAN KINOME

Jose Izarzugaza*, Angela del Pozo, Alfonso Valencia

* *Spanish National Cancer Research Institute, Spain*
email: jmgonzalez@cnio.es

Human Protein Kinases are involved in a wide variety of physiological functions. Most of the many mutations described in this protein family are tolerated without significant disruption of their structure or function. Interestingly, a number of them are associated to human diseases, including cancer, and deserve particular attention. Here we present the basis for the development of a computational method for the prediction of the impact of mutations in the function of protein kinases. The study was carried out in a set of 3492 well-characterized disease and neutral kinase-mutations extracted from Uniprot. We explored the significance of disease-associated kinase mutations in terms of sequence-derived characteristics at different levels, including: a) at the gene level, the membership to a Kinbase group and Gene Ontology terms. b) at the domain level, the occurrence of the mutation inside a PFAM domain, and c) at the residue level, several properties including amino acid types, functional annotations from Swissprot and FireDB, specificity-determining positions, etc. We analyzed the independent significance of these properties and their combination, with a Support Vector Machine (SVM). Interestingly, the family-specific features appear among the most discriminative information sources, which justifies the development of a kinase-specific predictor. Our study aims to broaden the knowledge on the mechanisms by which mutations in the human kinome contribute to disease with a particular focus in cancer. In addition, we discuss the benefits and pitfalls of using the information available for the development of a kinase-specific predictor with regard to other current prediction methods.

LARGE-SCALE COMPUTATIONAL IDENTIFICATION OF REGULATORY SNPS

Alberto Riva*

*University of Florida, USA
email: ariva@ufl.edu

rSNP-MAPPER is a web-based tool to identify SNPs that potentially affect a Transcription Factor Binding Site to a significant extent (regulatory SNPs, or rSNPs). rSNP-MAPPER builds on MAPPER, a previously developed application for the computational detection of TFBSs in DNA sequences. We have provided MAPPER with the ability to analyze two variants of the same sequence (each containing one allele of a SNPs), determining whether the substitution results in a significant change in the TFBS predictive score. The application's provides an intuitive and flexible interface. The user may search for potential rSNPs in the promoters of one or more genes, that can be specified as a list, or chosen from the members of a pathway. Alternatively, the user may specify a set of SNPs to be analyzed by uploading a list of SNP identifiers or providing the coordinates of a genomic region; rSNP-MAPPER will determine which SNPs lie within a TFBS and compute the corresponding score changes. Finally, the use can provide two alternative sequences (wildtype and mutant): the system will analyze them to determine the location of variants, to identify potential TFBSs, and to determine the effect of the variants on the TFBS scores. rSNP-MAPPER is optimized to efficiently perform all these operations on a large scale, allowing for the fast annotation of thousands of SNPs. We present the architecture of rSNP-MAPPER, and we describe its usage through several examples that demonstrate its ability to correctly identify previously known rSNPs, or to predict new ones with high confidence.

A LARGE SCALE ANALYSIS IN THE HUMAN PROTEOME DETECTS CORRELATION AMONG DISEASE ASSOCIATED MUTATIONS AND PERTURBATION OF PROTEIN STABILITY

Valentina Indio, Pier Luigi Martelli, Marco Vassura, Piero Fariselli, Rita Casadio*

*University of Bologna, Italy
email: casadio@biocomp.unibo.it

Technological advancements constantly increase the number of mutations that need annotation in translated regions of the human genome. Single residue mutations in proteins are known to affect protein stability and function. As a consequence they can be disease associated. Available computational methods starting from protein sequence/structure predict whether residue mutations are conducive to disease or alternatively to instability of the protein folded structure. However the relationship among stability changes in proteins and their involvement in human diseases still needs to be established. Here we try to rationalize in a nutshell the complexity of the question by generalizing over information already stored in public databases. For this we derive for each Single Aminoacid Polymorphysm (SAP) type the probability of being disease-related (Pd) and compute from thermodynamic data three indexes indicating the probability that it is conducive to decreasing (P-), increasing (P+) and perturbing the protein structure stability (Pp). Statistically validated analysis of the different P/Pd correlations indicates that Pd best correlates with Pp. Pp/Pd correlation values are as high as 0.49, and increase up 0.67 when data variability is taken into consideration. This is indicative of a medium/good correlation among Pd and Pp and corroborates the assumption that protein stability changes can be associated to disease at the proteome level. All the probabilities are listed in a feature table useful to label SAPs as disease/protein perturbation frequently or less frequently associated in the current data bases.

SNPS&GO: PREDICTING THE DELETERIOUS EFFECT OF HUMAN MUTATIONS USING FUNCTIONAL ANNOTATION

Emidio Capriotti*, Piero Fariselli,, Pier Luigi Martelli, Rita Casadio

*Stanford University, USA
email: emidio@stanford.edu

High-throughput data from large-scale sequencing and genotyping techniques allow to analyze a huge amount of genetic variation from whole human genome. Single Nucleotide Polymorphisms (SNPs), which are the main cause of human genome variability, can also be involved in the insurgence of many diseases. In particular missense SNPs, occurring in coding regions and causing single amino acid polymorphisms (SAPs), can affect protein function and lead to genetic pathologies. In this work, we present SNPs&GO (Calabrese at al., Human Mutation 2009), a new web server for the prediction of deleterious SAPs using protein functional annotation. We implemented two different SVM-based methods relying either on protein sequence or structure information. Both algorithms have been extensively tested on a large set of mutations extracted from SwissVar database (Mottaz et al., Bioinformatics 2010). Selecting a balanced dataset of SAPs, the sequence-based approach reaches 81% overall accuracy, 0.63 correlation coefficient and 0.89 area under the receiving operating characteristic curve (AUC). For the subset of mutations that can be mapped on protein structures known with atomic resolution (at the Protein Data Bank), the structure-based method results in 85% overall accuracy, a correlation coefficient of 0.70, and AUC of 0.92. In conclusion, SNPs&GO is a valuable tools that includes in unique framework information derived from protein sequence, structure, evolutionary profile, and protein function. In a recent study (Thusberg et al., Human Mutation 2011), SNPs&GO has been scored as one of the best algorithms for prediction of deleterious SAPs.
Avaliablity: http://snps.uib.es/snps-and-go

COMPARE H. SAPIENS + H. NEANDERTHALENSIS BY PREDICTING SNPS EFFECTS

Shaila Rössle* Dominik Achten, Martin Kircher, Janet Kelso, Svante Pääbo, Burkhard Rost

* Technical University Munich, Germany
email: roessle@rostlab.org

We analyzed features that are unique to human with respect to Neandertal.We started from a set of 78 nucleotide substitutions (nsSNPs) where modern humans are fixed for a derived state and where the Neandertal carry the ancestral state. We applied SNAP to these mutations in order to predict changes that affect protein function and infer phenotypic differences between humans and Neandertals. In particular, for each nsSNP encountered in 72 human, chimpanzee and Neandertal proteins, we predicted the effect of all non-native mutations. Our results are still very preliminary; we focus on very few cases of proteins for which the predicted sensitivity to change differed substantially, proteins that show high sequence identity between human and chimpanzee, and that SNP shows different functional effects among the organisms predicted by SNAP. Neandertal proteins with very high similarity between human and chimpanzee could show phenotypic differences. Sometimes resemble human proteins in terms of their sensitivity to mutations and sometimes they seem to be more chimpanzee proteins. For example SSH2, a protein phosphatase, differs from chimpanzee in its interaction hotspots. On the other hand ACCN4, an acid sensing ion channel implicated in synaptic transmission, pain perception as well as mechanoperception, by our analysis is more similar to the human protein. Our major challenge is getting further remain the incompleteness of Neandertal: 10% have been sequenced. This number implies that not a single Neanderthal protein is properly known. Should we use human to fill white spaces or chimpanzee? Or a hybrid? Next, we will address these.

THE FUNCTIONAL IMPORTANCE AND DETECTION OF REGULATORY SEQUENCE VARIANTS

Virginie Bernard*, Wyeth Wasserman, David Arenillas
*University of British Columbia, Canada
email: virginie@cmmt.ubc.ca

The convergence of high-throughput technologies for sequencing individual exomes and genomes and rapid advances in genome annotation are driving a neo-revolution in human genetics. This wave of family-based genetics analysis is revealing causal mutations responsible for striking phenotypes. By mapping the reads to the human genome reference and by searching for variations relative to the reference, a list of small nucleotide variations and structural variations is obtained. Analysis is required to reveal those variations most likely to contribute to a disease phenotype within a family. Existing software score the severity of changes that arise in protein encoding exons. However, most mutations within a family are situated in the 98% of the genome that controls the developmental and physiological profile of gene activity - the sequences that control when and where a gene will be active. Functional contributions of cis-regulatory sequence variations to human genetic disease are numerous. With full genome sequencing becoming accessible to medical researchers, the need to identify potential causal mutations in regulatory DNA is becoming imperative. We are implementing a software system to enable genetics researchers to characterize regulatory DNA changes within individual genome sequences. We are combining reference databases of known regulatory elements, experimental archives of protein-DNA interactions and computational predictions within an integrated analysis package. With our software, researchers will have greater capacity to identify variations potentially causal for disease. The poster introduces the challenges and approaches of regulatory sequence variation analysis.

LARGE-SCALE PROTEIN FLEXIBILITY ANALYSIS OF SINGLE NUCLEOTIDE POLYMORPHISMS, USING MOLECULAR DYNAMICS SIMULATIONS

Marc Offman*, Burkhard Rost
* Technical University of Munich, Germany
email: offman@rostlab.org

Proteins are intrinsically flexible molecules, thus function is often associated to flexibility. Experimental methods to determine protein flexibility are expensive and often time consuming. Over the past few years an efficient complementing method, molecular dynamics simulations, more and more proved to be a powerful tool to yield information on protein dynamics. We have recently proven that careful biology-driven MD simulations can be used to predict the impact of single amino acid mutations on protein flexibility and function, at a level of accuracy comparable to experimental techniques. The question remains whether it will be possible to fully automate this process in the context of a large-scale analysis, and to what extent additional structural information, beyond that derived by sequence analysis of single nucleotide polymorphisms (SNPs) only, is useful. For this we created several comprehensive datasets of non-synonymous SNPs mapped to high-resolution and above average quality crystal structures from the PDB. In the context of the European SCALALIFE project up to 28,000 different mutations found in 1,600 individual crystal structures are simulated in duplicates for 10 ns each, using the GROMACS package. A comprehensive analysis pipeline has been established, investigating protein flexibility and stability, alteration of hydrogen-bond networks, active site integrity, changes in global and local energy and other structural effects. This pipeline has previously been successfully applied in the context of clinically relevant proteins. The results of this study, the automatic protocol and set of analysis tools will help in the future to understand individual phenotypes in clinical contexts.

CLASSIFICATION OF MISMATCH REPAIR GENE MISSENSE VARIANTS

Heidi Ali*, Ayodeji Olatubosun,, Mauno Vihinen

** University of Tampere, Finland*
email: heidi.ali@uta.fi

Lynch syndrome accounts for approximately 2 to 5 % of colorectal cancers. The syndrome is caused by germline mutations in mismatch repair (MMR) genes, MLH1, MLH3, MSH2, MSH6, PMS1, PMS2 and TFGBR2. MMR is a DNA repair system that recognizes and repairs base-base mispairs and insertion-deletion loops arising in DNA replication and recombination. Thousands of MMR variants have been discovered, but their relevance to the cancer is usually unknown. Here, we utilize bioinformatics prediction methods to classify MMR variants. We identified from literature 168 functionally tested MMR missense variants of which 82 were pathogenic. The InSiGHT database for Lynch syndrome data contains over 600 variants with unknown effect.
We used Pathogenic-Or-Not-Pipeline http://bioinf.uta.fi/PON-P for the prediction and analysis of these variants. Since the performance of the individual predictors was not as good as we wanted, we developed a consensus predictor based on several tolerance prediction methods. With this predictor, we were able to classify over 200 previously unknown MMR missense variants as pathogenic or neutral. The results can be used to prioritize variants for further experimental validation and may help in the diagnosis of Lynch syndrome and other gastric cancers.

FUNCTIONAL ANNOTATION OF REGULATORY VARIANTS: A SYSTEMS BIOLOGY APPROACH TO TRANSLATIONAL BIOINFORMATICS

Konrad Karczewski*, Joel Dudley, Nicholas Tatonetti, Stephen Landt, Atul Butte, Michael Snyder, Russ Altman

** Stanford University, USA*
email: konradjkarczewski@gmail.com

Many genomic variants are discovered outside of genes, where their functional consequences are more difficult to characterize. However, as many of these variants are associated with disease, it is likely that they affect molecular physiology at the level of gene regulation. We investigate the role of variants in regulatory regions, on both transcription factor cooperativity as well as disease pathophysiology. First, we developed the Allele Binding Cooperativity (ABC) test and the ALPHABIT pipeline, which utilizes variation in transcription factor binding among individuals to discover combinations of factors and their targets. Second, we demonstrate a systematic approach to combine disease association, transcription factor binding, and gene expression data to assess the functional consequences of variants associated with hundreds of human diseases. In this systematic approach, we close a major loop in biological context-free association studies and assign putative function to many disease-associated SNPs. In this way, we apply findings from systems biology in a translational approach.

COMPUTATIONAL METHODS AS FIRST-PASS FILTER FOR MISSENSE MUTATIONS IN ATM GENE- A ROAD TOWARDS PHARMACOGENOMIC APPROACH

George Priya Doss *, Ujjwal Shah, Srajan Jain

** VIT University, India*

email: georgepriyadoss@vit.ac.in

With the recent availability of the complete genome sequence and the accumulation of variation data, determining the effects of missense variants will be the next challenge in mutation research. In addition to the molecular approaches, which are laborious and time-consuming, it is now possible to apply computational approaches to filter out mutations that are unlikely to affect protein function. Alternatively, bioinformatics approaches, based on the biochemical severity of the amino acid substitution, and the protein sequence and structural information, can offer a more feasible means for phenotype prediction. Deleterious missense mutations of ATM gene are accountable for various forms of cancer associated disease. Yet, distinguishing deleterious mutations of ATM from the massive number of non-functional variants that occur within a single genome is a considerable challenge. In this approach, we present the use of computational methods to explore the mutation-structure-function relationship. In other respects, we attempted these methods to work as first-pass filter to identify the missense mutations worth pursuing for further experimental research. This review surveys and compares variation databases and in silico prediction programs that assess the effects of deleterious functional variants on protein functions. We also introduce a combinatorial approach that uses machine learning algorithms to improve prediction performance.

PATHOGENIC-OR-NOT-PIPELINE (PON-P): INTEGRATION OF PREDICTORS FOR DISEASE RELEVANCE

Ayodeji Olatubosun, Jouni Väliaho, Mauno Vihinen*

** University of Tampere, Finland*

email: mauno.vihinen@uta.fi

This work deals with the prediction of whether a non-synonymous single nucleotide polymorphism leads to disease or not. This is an important scientific problem, with potential applications in such areas as molecular diagnosis, prioritization of experiments and screening of variations. This work utilizes a unique approach in an effort to improve on the current performance limitations of predictors in this problem domain. The importance of this work to this particular conference is further highlighted by the fact that the SNP-special interest group one-day meeting is organized specifically to address the same problem dealt with in this work, which will also be the focus of the second special session in this meeting. The results of this work, and their implications would be of high interest to researchers in many fields.

WHOLE GENOME SEQUENCING: FROM ENU-INDUCED MUTATIONS TO MOUSE MODELS OF HUMAN DISEASE

Michelle Simon*, Simon Greenaway,
Paul Denny, Paul Potter, Anne-Marie Mallon,
Steve Brown

* *Medical Research Council, UK*
email: m.simon@har.mrc.ac.uk

Phenotype-driven screens after chemical mutagenesis of males with N-ethyl-N-nitrosourea (ENU) at MRC Harwell have been incredibly productive. Nevertheless, identification of the causative mutations by conventional linkage analysis and sequencing of genes in the minimal genetic interval remains a bottleneck. We have been using next generation sequencing (NGS), using the Illumina Genome Analyser 2x platform, to accelerate the process of mutation detection. A custom sequence analysis pipeline has been developed to capture and analyse causative mutations. The pipeline is based on existing packages (i.e. Bowtie, Samtools, CASAVA, nFold3, etc.) and custom developed components. The first part of the pipeline is used to align reads to the C57BL/6J reference sequence, automatically identify unique variants, populate a custom sequence database and identify low and high confidence single nucleotide polymorphisms (SNPs). The second part includes assessment of the impact of putative mutations on splicing efficiency, predicted protein structure and phenotype. The EuroPhenome database at MRC Harwell captures high-throughput phenotypic data from a number of projects such as EUMODIC and in the future from the Harwell Ageing Screen. The NGS data from our pipeline will be integrated with phenotyping data in EuroPhenome and enable us to investigate the impact on mouse phenotypes and ultimately predict the likelihood of this as a model for human disease. We will present the re-sequencing pipeline, the identification of unique variants and the possible impact of any SNPs on phenotype.

TESTING FOR JOINT ASSOCIATION OF ALL SNP PAIRS

Ronald Schuyler*, Lawrance Hunter

* *University of Colorado Denver USA*
email: ron.schuyler@gmail.com

Despite the success of genome-wide association studies (GWAS) in providing insight into mechanisms of disease, the associated loci usually account for only a fraction of the expected heritability of each condition studied. This implies that more may be learned from GWAS data by going beyond the one-SNP-at-a-time association approach. In studying complex traits, it is useful to look for associations with pairs of loci. The standard logistic regression test for joint effects requires iterative methods for determining maximum likelihood estimates (MLEs), which makes testing all possible pairwise combinations from common high-throughput genotyping platforms extremely computationally demanding, and limits the wider application of this approach. Loglinear categorical data analysis methods with closed-form solutions for MLEs are well known and have recently been applied to GWAS data, reducing computation time and making it feasible to test all locus pairs. We have proposed a refinement to this method which reduces the number of computations by nearly two thirds. We used likelihood ratio tests of loglinear models with a step-wise model selection procedure to test all 150 billion possible two-locus pairs of a 550k SNP study for joint association with generalized vitiligo. Using a stringent multiple testing adjustment, we detected a small number of significant pairs where the two SNPs of every pair detected are in close proximity to one another. SNPs in most regions detected showed no trend toward significance in single locus tests, but these loci have clear biological relevance to the condition studied.

IMPROVING THE DETECTION OF DELETERIOUS MUTATIONS INTEGRATING THE PREDICTIONS FOUR WELL-TESTED METHODS

Emidio Capriotti*, Yana Bromberg, Russ Altman

** Stanford University, USA*

email: emidio@stanford.edu

In the past few years the number of human genetic variations deposited in the web available databases has been increasing exponentially. The last version of dbSNPs (build 132) contains about 20 million validated Single Nucleotide Polymorphisms (SNPs). SNPs make up most of human variation and are often the primary causes of disease. The coding region non-synonymous SNPs (SAPs) result in amino acid changes and may affect protein function causing severe genetic diseases. Although several methods for the detection of SAPs have been implemented, the increasing amount of annotated data is now offering the opportunity to develop more accurate algorithms. Here we present an approach for the prediction of the effect that integrates four methods including PANTHER, PhD-SNP, SIFT and SNAP. We first tested the accuracy of each method using a dataset of 35,986 annotated mutations from 2,269 proteins extracted in October 2009 from SwissVar database. The four methods reached overall accuracies ranging between 64% and 76% and a correlation coefficient from 0.38 to 0.53. We then developed an SVM-based approach that takes as input a ten elements vector derived for the output of the 4 methods. When tested using a cross-validation procedure, the integrated method reaches 80% overall accuracy and 0.60 correlation coefficient resulting in 4% higher accuracy and 0.07 higher correlation coefficient with respect to the best method.

Availability: http://snps.uib.es/meta-snp

PROTEIN-LEVEL EFFECTS OF MUTATIONS IN OVARIAN CANCER, ACUTE MYELOID LEUKEMIA AND GLIOBLASTOMA MULTIFORME

Janita Thusberg*, Charles Vaske, Zack Sanborn, Joshua Stuart, Christopher Benz, David Haussler, Sean Mooney

** Buck Institute for Research on Aging, USA*

email: jthusberg@buckinstitute.org

The heterogeneity of mutation profiles in cancer patients calls for detailed annotation of the putative downstream effects of mutations. A subset of the somatic mutations is expected to function as drivers, and identifying variants responsible for tumor progression among dozens, or even hundreds within a patient, is not a trivial task. Protein-level annotation of mutations may reveal functions beyond the analysis of genes enriched in mutations, providing molecular level explanations for the roles of missense mutations in cancer progression. By elucidating protein-level effects of missense mutations common patterns among cancers can be found. Coding variants identified from exon-capture and whole genome sequencing of tumor samples have been bioinformatically characterized by a suite of applications to identify variants likely to disrupt molecular function or clinical phenotype and hypothesize their molecular effects. Based on a training set of disease causing mutations and neutral polymorphisms, the program MutPred utilizes a Random Forest method to predict functional variants using proteomic, genomic and bioinformatic attributes. Since these attributes are based on known and predicted protein functional sites, the disrupted function can be quantitatively hypothesized. The MutPred scores and functional attributes identify variants likely to drive tumor progression and are also utilized further with other genomic data in pathway analysis. PARADIGM is a probabilistic graphical model that integrates genomic and functional genomic data to infer tumor-specific alterations in gene activity in the context of known gene pathways. The protein-level interpretations of mutations are used to extend the PARADIGM model of gene disruptions beyond deletions and amplifications.

FUNCTIONAL PROFILING OF PHARMACOGENETIC NON-SYNONYMOUS SNPS

Janita Thusberg, Emidio Capriotti, Jim Auer, Sean Mooney*

* *Buck Institute for Research on Aging, USA*
email: smooney@buckinstitute.org

Bioinformatic study of the effects of disease-related SNPs is well established and several tools for annotation and prediction of the effects of these variants have been developed. Due to large-scale genotyping and sequencing efforts, increasing amounts of knowledge about genetic variants associated with diseases, complex phenotypes as well as drug response and metabolism is accumulating in the literature and databases, that will be invaluable data for personal genetics applications and also for clinical setting.

Little is known about the nature of pharmacogenetics variants as compared to disease-causing mutations and neutral polymorphisms, and whether the same methods can be utilized in studying and predicting disease-causing variants and pharmacogenetic SNPs. We have annotated the protein level consequences of pharmacodynamic and pharmacokinetic variants in PharmGKB by bioinformatics methods and elucidated features that differentiate a pharmacogenetic variant from other types of genetic variants. We analyzed a set of 352 SNPs from 222 proteins that have been annotated in the SwissVar database. We found that 92% of them were annotated as neutral polymorphisms and only 74% of them were correctly annotated by mutation prediction algorithms. This suggests that about one over four mutations could have some functional effect. In the near future, the results of this analysis are aiming to provide the characteristic features that can be used to define a pharmacogenetic variant fingerprint, further utilized in the development of methods for discovering new variants with putative pharmacogenetic effects, and to hypothesize new biomarkers for predicting drug response.

PREDICTING THE FUNCTIONAL IMPACT OF PROTEIN MUTATIONS: APPLICATION TO CANCER GENOMICS

Boris Reva*, Yevgeniy Antipin, Chris Sander

* *Memorial Sloan-Kettering Cancer Center, USA*
email: borisr@mskcc.org

As large scale re-sequencing of genomes reveals many protein mutations, especially in human cancer tissues, prediction of their likely functional impact becomes an important practical goal. Here, we introduce a new functional impact score (FIS) for amino acid residue changes using conservation patterns. The evolutionary information in these patterns is derived from aligned families and sub-families of sequence homologs within and between species using a combinatorial entropy formalism. We tested the score on a large set of human protein mutations for its ability to separate disease-associated variants (~19,200), assumed to be strongly functional, from common polymorphisms (~35,600), assumed to be weakly functional, and obtained an area under the receiver-operating-characteristic curve of ~0.86. From analysis of ~10,000 cancer mutations of COSMIC, we conclude that recurrent mutations, mutations in multiply mutated genes and mutations annotated as cancer genes tend to have significantly higher functional impact scores than control sets. We report a ranked list of ~1000 top human cancer genes frequently mutated in one or more cancer types; and, an additional ~1000 candidate cancer genes with rare but likely functional mutations. In addition, we estimate that ~5% of cancer-relevant mutations involve switch of function, rather than simply loss or gain of function. The computational protocol is implemented as a public server: http://mutationassessor.org. The server provides links to multiple sequence alignment and 3D structures, to various biological and cancer annotations. The service is built to process output of sequencing machines, and it is capable of quickly processing thousands of variants (through WEBAPI).

SNPEFFECT 4.0: MOLECULAR AND STRUCTURAL PHENOTYPING OF HUMAN SNPS AND DISEASE MUTATIONS

Greet De Baets*, Joost Schymkowitz, Fredric Russeau

** Free University Brussels, Belgium*
email: greet.debaets@switch.vib-vub.be

Single nucleotide polymorphisms (SNPs) are, together with copy number variation, the primary source of variation in the human genome and are associated with altered response to drug treatment, susceptibility to disease, and other phenotypic variation. Linking structural effects of non-synonymous SNPs to functional outcomes is a major issue in structural bioinformatics, and many tools and studies have shown that specific structural properties such as stability and residue burial can be used to distinguish neutral variations and disease associated mutations. The SNPeffect database uses sequence- and structure-based bioinformatics tools to predict the effect on the molecular phenotype of proteins. It integrates Tango (an aggregation predictor, http://tango.crg.es/); Waltz (a predictor of amyloid forming sequences, http://waltz.switchlab.org/); Limbo (a predictor for chaperone specificity, http://limbo.switchlab.org/); and FoldX (http://foldx.switchlab.org/) that reports the $\Delta\Delta G$, the change in free energy upon mutation. In that way, FoldX predicts the effect of SNPs in two categories of functional properties: (1) structural and thermodynamic properties affecting protein dynamics and stability and (2) the integrity of functional and binding sites. The database already contains the annotations for the UniProt set of human disease and polymorphism mutations, but users can also submit their own set of mutations for analysis.

ANALYSIS OF STRUCTURAL AND FUNCTIONAL IMPACTS OF BRUTON TYROSINE KINASE MUTATIONS

Jouni Väliaho*, Mauno Vihinen

** University of Tanpere, Finland*
email: jouni.valiaho@uta.fi

X-linked agammaglobulinemia (XLA) is caused by mutations in the gene encoding Bruton tyrosine kinase (BTK). XLA patients have a decreased number of mature B cells and a lack of all immunoglobulin isotypes, resulting in susceptibility to severe bacterial infections. We are collecting XLA-causing mutations are collected in a mutation database (BTKbase), which is available at http://bioinf.uta.fi/BTKbase/. The database contains 1155 cases coming from all around the world. Btk protein consist of five distinct structural domains, from the N-terminus: pleckstrin homology (PH), Tec homology (TH), Src homology 3 (SH3), SH2, and the catalytic kinase domain (TK).

We did a detailed analysis of XLA causing missense mutations in Btk tyrosine kinase domain by using numerous methods for predicting the effects of amino acid substitutions. We have utilized the available 3D structures of Btk kinase domain for studies of the contacts between residues, their implication for the stability of the protein, and the effects of the introduced residues. Investigations of steric and stereochemical consequences of substitutions provide insights on the molecular fit of the introduced residue. Mutations that change the electrostatic surface potential of a protein have wide-ranging effects. Analyses of the effects of mutations on interactions with ligands and partners have been performed for elucidation of functional mutations. Detailed analyses of the variations have allowed us to provide tentative explanation for all the 478 unique variations found in XLA patients.

A NOVEL COMBINED SCORE PREDICTS THE EFFECT OF NON-SYNONYMOUS SNPS ROBUSTLY

Margarida Lopes*, Chris Joyce, Jennifer Asimit, Eleftheria Zeggini

*Wellcome Trust Sanger, UK
email: ml10@sanger.ac.uk

Next generation sequencing has opened the possibility of large-scale sequence-based disease association studies. A major challenge in interpreting whole-exome data is predicting which of the discovered variants are deleterious or neutral. To address this question, in the absence of further experimental research, several functional-annotation tools focusing on the analysis of non-synonymous coding SNPs (nsSNPs) have been implemented. We developed a meta-predictor which combines information from three bioinformatics tools: PolyPhen-2, SIFT and PANTHER, in order to improve the prediction of the effect of nsSNPs. We use a weighted-Z method, which combines the probability (P(a,i)) of the substituted amino acid (a) occurring at each position (i) from protein alignments. The weighting algorithm is linearly related to Genomic Evolutionary Rate Profiling (GERP) scores and inversely proportional to P(a,i), overweighting amino acid substitutions that lie in evolutionary conserved regions across multiple species . For validation purposes we selected a set of 2,944 monogenic disease-causing nsSNPs and 64,917 (presumed) neutral nsSNPs to be used as our positive and negative controls, respectively. This meta-predictor is expected to perform better than individual annotation tools, thus facilitating the analysis and interpretation of next-generation association studies.

PROTEIN STRUCTURE ANALYSIS OF MUTATIONS CAUSING INHERITABLE DISEASES. AN E-SCIENCE APPROACH WITH LIFE SCIENTIST FRIENDLY INTERFACES.

Gert Vriend*, Hanka Venselaar*

*Radboud University, Netherlands
email: vriend@cmbi.ru.nl, h.venselaar@cmbi.ru.nl

Many newly detected point mutations are located in protein-coding regions of the human genome. Knowledge of their effects on the protein's 3D structure provides insight into the protein's mechanism, can aid the design of further experiments, and eventually can lead to the development of new medicines and diagnostic tools. In this article we describe HOPE, a fully automatic program that analyzes the structural and functional effects of point mutations. HOPE collects information from a wide range of information sources including calculations on the 3D coordinates of the protein by using WHAT IF Web services, sequence annotations from the UniProt database, and predictions by DAS services. Homology models are built with YASARA. Data is stored in a database and used in a decision scheme to identify the effects of a mutation on the protein's 3D structure and function. HOPE builds a report with text, figures, and animations that is easy to use and understandable for (bio) medical researchers. We tested HOPE by comparing its output to the results of manually performed projects. In all straightforward cases HOPE performed similar to a trained bioinformatician. The use of 3D structures helps optimize the results in terms of reliability and details. HOPE's results are easy to understand and are presented in a way that is attractive for researchers without an extensive bioinformatics background.

CONSIDERING PROTEIN CONFORMATIONAL DIVERSITY IMPROVES DISEASE ASSOCIATED MUTATIONS PREDICTION

Ezequiel Juritz*, Pier Luigi Martelli, Maria Silvina Fornasari, Rita Casadio, Gustavo Parisi

** Universidad Nacional de Quilmes, Argentina*
email: eze1982ar@gmail.com

Great efforts have been made in the last years to predict the effect of point mutations in protein native conformation in relation to disease phenotypes. Most available methods suited to predict/compute the effect of a single side chain substitution on protein stability take as input the protein structure. However, protein function highly depends on the presence of an ensemble of different conformations describing the functional native state. Here we explore the effect of protein structural conformational diversity on predicting whether a mutation that is annotated as disease associated is also affecting the protein stability. For this purpose, we adopted a data base of protein structures and their associated conformers as derived from CATH [PCDB;http://www.pcdb.unq.edu.ar]. 312 proteins with associated conformers are endowed with 5642 disease associated and 235 polymorphic mutations, respectively, as annotated in UniProtKB(September 2010). For each mutation in each protein we computed the Gibbs free energy variation (DDG) using I-Mutant[http://gpcr.biocomp.unibo.it/~emidio/I-Mutant/I-Mutant.htm] and Fold-X[http://foldx.crg.es] and compared the values obtained with a single structure or with a compilation of different conformers per protein. We found that the explicit consideration of protein conformational diversity increases the predictive power of DDG measures as derived from ROC plots and accuracy calculations. These results are in accordance with the fact that a conformational ensemble description of a protein better describes the protein function and increase the reliability to explore the possible causes of function lost and disease phenotypic occurrence.

TOWARDS LINKED OPEN GENE MUTATIONS DATA

Achille Zappa, Andrea Splendian, Paolo Romano*

**National Institute for Cancer Research, Italy*
email: paolo.romano@istge.it

Semantic Web technologies are enough mature to offer a viable solution for data integration. A requirement for this is the conversion into RDF of data stored in relational databases. Although the human variation analysis is now one of the biggest issues, there is no mutation data available in RDF. In this poster, we present a first prototype of a server implementing an RDF representation of the IARC TP53 Mutation Database. This prototype server was developed with the aim of studying all issues related to the publication of linked mutation data. It was developed by using the D2RQ platform. Automatic mappings were first generated. Later a fine tuned, manual revision of mappings was carried out in order to incorporate proper relationships making reference to ontologies widely adopted in the field, e.g., bibo (bibliographic ontology), and to link to external systems by the use of standardized URIs. A prototype D2R server is now available on-line at http://ml370.istge.it:7777/. It includes somatic mutations, i.e., data on observed mutations, gene variations, summarizing effects of known mutations, and related bibliographic references. An HTML view and a SPARQL endpoint are available starting from this address. Linked Data views can be obtained from other sites. This prototype demonstrates that an RDF representation of mutation data can already be easily set up. The main difficulty lies, as usual, on the identification of a shared, semantically meaningful, ontology-based representation of variation information. A revised version of the prototype including more shared concepts and the full IARC database is under development.

HUNTING CANCER PREDISPOSITION GENES BY THE WHOLE EXOME SEQUENCING IN RELATIVES FROM AN AFFECTED FAMILY: SNPS IN FOCUS

Tatiana Popova*, Virginie Jacquemin, Severine Lair, Romain Daveau, Emmanuel Barillot, Dominique Stoppa-Lyonnet, Marc-Henri Stern*

**Institut Curie, France*

email: tatiana.popova@curie.fr, marc-henri.stern@curie.fr

A whole exome sequencing of constitutional DNA from two relatives of a family with high incidence of cancer is presented. Based on this example we consider a potential of the Next Generation Sequencing (NGS) study for identifying familial cancer predisposition gene. We present the data analysis workflow with a special emphasis on the SNPs detection, inference of identical by descent chromosomal regions based on the detected SNPs and analysis of possible sources of the false positive SNP calls.

SNP FUNCTION PORTAL VERSION 2: NOVEL VARIANTS, BETTER USABILITY AND OPEN ARCHITECTURE

Josh Buckner, Weijian Xuan, Manhong Dai, Justin Wilson, Brian Athey, Fan Meng*

**University of Michigan. USA*

email: mengf@umich.edu

We overhauled the SNP Function Portal to provide extensive function annotation for novel variants identified in deep sequencing data analysis, better usability of search results and open architecture for users to load and share their custom annotations.

ACKNOWLEDGMENTS

The SNP-SIG meeting organizers would like to acknowledge the highlight and keynote speakers:

- Steven Brenner, UC Berkeley, USA
- Atul J. Butte, Stanford University, USA
- Burkhard Rost, Technical University Munich. Germany
- Mauno Vihinen, Tampere University, Finland

and the following scientists involved in the reviewing process and roundtable organizers:

- Leonardo Arbiza, Cornell University, USA
- Christopher J Baker, University of New Brunswick, Canada
- Dario Boffelli, UC Merced, USA
- William S. Bush, Vanderbilt University, USA
- Rong Chen, Stanford University, USA
- Jianlin Cheng, University of Missouri, USA
- Francisco Xavier De La Cruz, IBMB, Spain
- Hernan Dopazo, CIPF, Spain
- Andre Franke, Christian Albrechts University, Germany
- Liang-Tsung Huang, Mingdao University, Taiwan
- Andrew Johnson, NIH, USA
- Maricel Kann, University of Maryland, USA
- Vidhya G Krishnan, A*STAR, Singapore
- Peter Kang, Stanford University, USA
- Adam Kowalczyk, University of Melbourne, Australia
- Phil Hyoun Lee, Queen's University, Canada.
- Jing Li, Case Western Reserve University, USA
- Marc A. Marti-Renom, CIPF, Spain
- Sean Mooney, Buck Institute, USA
- Matthew Mort, Cardiff University, UK
- Alessio Naccarati, Institute of Experimental Medicine, Czech Republic
- Pauline C Ng, A*STAR, Singapore
- Yanay Ofran, Bar Ilan University, Israel
- Gaurav Pandey, UC Berkeley, USA
- Yue Peng, Genentech, USA
- Predrag Radivojac, Indiana University, USA
- Lipika Ray, University of Maryland, USA
- Susanna Repo, UC Berkeley, USA
- Joke Reumers, Free University of Brussels, Belgium
- Joost Schymkowitz, Free University of Brussels, Belgium
- Paul Thomas, University of Southern California, USA
- Janita Thusberg, Buck Institute, USA
- Ali Torkamani, Scipps, USA
- Mauno Vihinen, Tampere University, Finland
- Dennis P Wall, Harvard Medical School, USA
- Zemin Zhang, Genentech, USA
- Yiqiang Zhao, Buck Institute, USA

The organizers also acknowledge **BIOBASE International** (www.biobase-international.com) for its financial support
.

AUTHOR INDEX