Breadth-first Search Based Approach to Enumerating Chemical Compounds Containing Outerplanar Fused Benzene Ring Substructures

Jina Jindalertudomdee

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

Structure enumeration is a problem of generating all non-redundant chemical compounds based on a given constraint, such as a chemical formula. It is important in chemoinformatics since it appears as a subproblem of several critical problems, such as drug discovery and structure elucidation. Until now, various algorithms have been proposed to solve the structure-enumeration problem. With the goal of enumerating all possible structures, some tools require impractical computation time for a moderate number of atoms in the input chemical formula. Therefore, alternative tools that can be executed in a reasonable period of time, but have the restriction of the structure of enumerated compounds, were developed. For example, a combination of *BfsSimEnum* and *BfsMulEnum*[1] can enumerate acyclic compounds efficiently using a tree structure to represent a chemical compound, where nodes and edges are labeled by atom types and bond multiplicities, respectively.

This work presents a novel algorithm *BfsFusedBenzeneEnum*, the extension of BfsSimEnum and BfsMulEnum, for the enumeration of chemical compounds containing no cyclic substructures except for outerplanar fused benzene ring substructures. In BfsFusedBenzeneEnum, a tree structure is used to represent a chemical compound with outerplanar fused benzene ring substructures by defining a special atom type, called *b*, and a special kind of bond, called *a merge bond*, to represent a benzene ring and a fused bond between two benzene rings, respectively. A node labeled with an atom type *b* has an additional attribute called *a carbon position list*, to keep information about which carbon atoms in the corresponding benzene ring bond with which adjacent nodes.

We evaluated the accuracy and efficiency by comparing the number of enumerated structures and computation time of the proposed algorithm with those of MOLGEN[2], a well-known commercial structure generator representing a chemical compound by a graph. The results show that the number of enumerated structures was the same between both of them, while computation time of our algorithm was significantly less than that of MOLGEN. The main reason for this speed-up is that the enumeration of tree structures is less complicated than the enumeration of graphs. Another reason is the decrease of the number of nodes during the enumeration because a benzene ring is represented by a single node instead of six carbon nodes.

[1] Y. Zhao, M. Hayashida, J. Jindalertudomdee, H. Nagamochi, and T. Akutsu. Breadth-first search approach to enumeration of tree-like chemical compounds. *Journal of Bioinformatics and Computational Biology*, 11(6), 2013.
[2] R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rucker, and A. Wassermann. Molgen 5.0, a molecular structure generator. *Bentham Science Publishers Ltd.,* 2012.

Protein Secondary Structure Prediction: Raising the Bar

Gerrit Korff and Ernst Walter Knapp

Macromolecular Modelling Group, Free University Berlin, Germany

Knowledge of a protein's structure is key to understand its function. Advances in sequencing technologies have widened the gap between known protein sequences and structures. Computational methods provide a means to lessen this gap, by predicting the 3D structure of a protein from its sequence. For this purpose, determining the local order of a protein, i.e. the protein secondary structure, is an important step toward complete knowledge of protein structure. Based on the secondary structure, homologous proteins can be found. Hence, protein secondary structure prediction is crucial to structurally characterize a protein and to predict its 3D structure.

*SPARROW$^+$, a new secondary-structure prediction method, has achieved a higher prediction quality than any other currently popular method. In an eightfold cross-validation test, the Q_3-accuracy was determined to be 83.8 % and the generalized MCC 0.75. Hence, *SPARROW$^+$ outperforms competitors by about 2 %. The outstanding feature of *SPARROW$^+$ is its novel vector-valued classifier, which has proven to be a powerful alternative to the commonly used artificial neural networks. In addition, based on a prediction confidence measure, *SPARROW$^+$ also offers a residue-specific Q_3-accuracy estimation.

Khoesan Bushmen Display Distinctive DNA Methylation Landscapes

Brenna A. LaBarre[1,2], Vanessa M.Hayes[3], and Laura L. Elnitski[2]

[1]Bioinformatics Graduate Program, Boston University, Boston, USA
[2]Genome Functional Analysis Section, Translational and Functional Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Rockville, USA
[3]Human Comparative and Prostate Cancer Genomics Lab, Garvan Institute of Medical Research, Sydney, Australia

Epigenetics, or changes "on top of" the genetic code, can provide added insight to genomic programing. Direct methylation of DNA at CpG dinucleotides can act as a signal for cellular machinery, or reflect changing cellular conditions. Some DNA methylation patterns are inherited, but others can arise from extracellular signals. Unforgiving environments, like that of the Kalahari Desert in southern Africa, can mold the methylation landscapes of the residents, such as the hunter/gatherer Khoesan Bushmen, in a way that resembles that of our early human ancestors. One expects the patterns seen in these people to differ from those seen in people with more modernized lifestyles because of methylomic stressors like pollution and medical intervention. Using Illumina 450k Human Methylation arrays we can assess the methylation signal at more than 485k CpG locations across the genome. We compared blood-derived DNA samples from Bushmen to samples from other groups living in industrialized areas throughout southern Africa, as well as to those from some Europeans and publicly available samples from Dutch, Caucasian, Japanese and African-American subjects. This revealed differentially methylated loci in the Bushmen that are being analyzed for potential functional implications. Based on our samples and analysis, there are more than 12,000 methylation loci that differ statistically between the Bushmen and the control samples. More than 500 probes have been identified as significant and potentially biologically relevant -- and not known to be C-to-T SNP loci. These results so far show identifiable differences in methylomes from Bushmen and non-Bushmen. The differences, especially those that influence metabolic processes and the immune system, may derive from the environmental adaptations of these people in the absence of conveniences of developed lifestyles.

Stabilization of Rafts in Hepatocyte Canalicular Membrane through Proteins

Johannes Eckstein

Institute of Biochemistry, Charité – University Medicine Berlin, Germany

We developed a mathematical model of the self-organization of cellular membranes into different domains. The model includes lipids that are freely diffusing along the membrane and proteins fixedly attached to the membrane. The diffusion of the lipids is modeled as a random movement on a triangular lattice governed by nearest neighbor interaction energies. Phase separation into liquid-ordered (Lo) and liquid-disordered (Ld) domains is modeled by assigning two alternative ordering states to each lipid species and minimizing the nearest-neighbor ordering energies. Parameterization of the model was performed such that experimentally determined diffusion rates and phases in ternary lipid mixtures of model membranes were correctly recapitulated. The model consists of three major membrane lipids phosphatidylcholine (PC), sphingomyelin (SM) and cholesterol (CH). A phase separation into an Ld domain enriched in PC and an Lo domain enriched in SM and CH occurs. These lipid domains form as maze-like structures that keep growing over time until complete demixing occurs. The model also includes two different protein species with an affinity to either the lipids of the Lo or the Ld domain representing the proteins resident in the respective membrane domains. The proteins are able to stop the lipid domains from continuously growing to sizes that would make them vulnerable to detergents such as bile salts. The stabilizing effect of the proteins on the lipid-domain structures can prevent the membrane from bile-salt-induced damage.

Diversity of Marine Giant DNA Viruses

Tomoko Mihara[1], Koyano, H.[2], Goto, S.[1], Ogata, H.[1]

[1]*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan*
[2]*Laboratory of Biostatistics and Bioinformatics, iACT, Kyoto University, Japan*

Background: Since the first giant virus was discovered in 2003[1], genomic data of giant viruses (mostly of the Megaviridae family) has been increasing, but we still lack enough amount of representative genomic data of this group of viruses to draw a full picture of their genetic diversity. They were found to be abundant in the ocean[2]. In this study, using marine metagenomics data that contains a large number of Megaviridae sequences, we attempted to assess the diversity of Megaviridae in comparison with those of bacteria and archaea that inhabit the same environments. We used RNA polymerase genes, which are encoded in both cellular organisms and Megaviridae, as suitable markers for comparative analyses of the sequence diversity between cellular organisms and Megaviridae[3].

Method: First, we retrieved RNA polymerase (RPol) protein sequences from the UniProt and RefSeq viral databases by hmmsearch using an existing HMM profile of RPol subunit β and β′ sequences and recreated HMM profiles with these reference database sequences for each subgroup of megavirus proteins. We then performed hmmsearch using these new HMM profiles against metagenome data sets including those from the *Tara* Ocean project (http://www.embl.de/taraoceans/) and obtained RPol candidate sequences. Second, by using a phylogenetic placement software package, pplacer[4], we classified and taxonomically annotated the candidate sequences based on the phylogenetic positions where the sequences were placed. We obtained 17,485 Megaviridae, 93,175 bacterial, 10,592 archaeal and about 2000 eukaryotic/viral RPol sequences. We used only Megaviridae, bacterial and archaeal sequences as they were abundant enough for comparative purpose. Third, we aligned the RPol sequences on an HMM profile created based on Megaviridae, bacteria and archaea reference sequences. The alignment length was 1147 residues. Using this alignment, we measured the richness (i.e., number of sequence clusters) and phylogenetic diversity of Megaviridae and prokaryotic sequences. The analyses were performed using a sliding window of 100 amino acid residues with the step size of 10 amino acids, in order to cope with the fragmented nature of the metagenomics RPol sequences. Phylogenetic diversity was calculated as the sum of total branch lengths on a phylogenetic tree. Sequence richness was estimated by clustering sequences according to their sequence similarities with each other.

Result: Bacteria exhibited the highest number of sequences among the analyzed organismal groups, and showed the highest diversity as expected from their high abundance in the sea. However, when we control the number of sequences to be the same between organism groups by re-sampling, Megaviridae sequences always showed larger phylogenetic diversity scores than bacterial and archaeal sequences. We also found that bacterial sequences showed larger numbers of clusters than others at any positions of the alignment, but after resampling, Megaviridae sequences always showed significantly larger numbers of clusters than bacteria and archaea. In summary, our results indicate that the extant lineages of Megaviridae are more diverse than those of bacteria and archaea.

Reference:

[1]La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D. A giant virus in amoebae. *Science* 2003 **299** (5615): (2003)

[2]Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, Ferrera I, Sarmento H, Villar E, Lima-Mendez G et al. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J* 2013 **7**(9): 1678-1695.(2013)

[3]Sharma V, Colson P, Giorgi R, Pontarotti P, Raoult D. DNA-dependent RNA polymerase detects hidden giant viruses in published databanks. *Genome Biology and Evolution* 2014 **6**(7): 1603-1610.

[4]Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 2010
11:538.

Chromatin State Patterns at Sex-Biased Transcription Factor Binding Sites Provide Clues to the Mechanisms of Sex-Biased Hepatic Transcriptional Regulation

Gracia M. Bonilla and David J. Waxman

Bioinformatics Graduate Program and Department of Biology, Boston University, Boston, USA

Sex-biased patterns of gene expression in mammalian liver are widespread and affect numerous biological processes, leading to sex differences in metabolism and disease risk. The sex-differential stimulation of hepatocytes by male versus female plasma patterns of growth hormone (GH) dictates liver sex differences by regulation at the transcriptional level via the combinatorial interactions of several GH-responsive liver transcription factors, including HNF6, whose sex-biased binding to liver chromatin correlates strongly with the sex-biased expression of neighboring genes. The sex bias of HNF6 binding is enriched at regions of sex-differential chromatin accessibility (DNase hypersensitive regions, DHSs); however, many sex-biased HNF6 binding events occur at sites where chromatin is equally accessible in male and female liver. We hypothesize that in addition to sex-biased chromatin accessibility, several other factors, including sex-biased histone marks and sequence preference of sex-biased cofactors, determine the sex-dependent binding of HNF6. To investigate this proposal, we clustered sex-biased HNF6 binding sites according to the sex bias of their chromatin state, previously established for male and female mouse liver using genome-wide DHS analysis and six histone marks. We identified sex-biased HNF6 binding sites in the same chromatin state, and HNF6 binding sites in different chromatin states between male and female liver. Examples include: (1) female-biased HNF6 binding at sites located within enhancer states in female liver but at genomic regions in an inactive chromatin state in male liver, and correspondingly for male-biased HNF6 binding sites, indicating that sex-differential chromatin states can be a determinant of sex-differential HNF6 binding; and (2) female-biased HNF6 binding at sites that do not show sex-differences in chromatin accessibility and are in the same chromatin state (enhancer state) in male and female liver, and correspondingly for male-biased HNF6 binding sites, indicating that factors other than accessibility to chromatin and local chromatin state may confer sex bias in HNF6 binding. Sequence motifs differentially enriched near these HNF6 binding sites identified potential cofactors of HNF6, which could contribute to the sex bias of HNF6 binding at genomic regions in a sex-independent chromatin state. In doing these analyses, we developed a tool that automates the clustering of genomic regions by the patterns of differential regulation of their local chromatin environment, and that facilitates the integration of these regions with genome-wide maps of chromatin accessibility, histone modifications, transcription factor binding motifs, and gene expression. This tool can be used generally to integrate multiple genomic features regulated by a common underlying mechanism and study their function. Supported in part by NIH grant R01-DK33765 (to DJW).

Bayesian Method for HLA Genotyping from Whole-Genome Sequencing Data

Shuto Hayashi

Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan

Human leukocyte antigen (HLA) genes encode proteins used by the immune system to distinguish between “self” and “nonself”. Because of their polymorphism and similarity, it is difficult to determine HLA genotypes from whole genome sequencing (WGS) data. Here we present a Bayesian model for accurate HLA genotyping from WGS data. Our method achieved more than 95% accuracy for HLA-A, HLA-B, and HLA-C.

Cataloging VNTRs in the Human Genome

Yozen Hernandez, Yevgeniy Gelfand, Gary Benson

Bioinformatics Graduate Program, Boston University, Boston, USA

Variable Number Tandem Repeats (VNTRs), repetitive sequences of DNA that differ in copy number among members of a species, have substantial clinical importance with several diseases having been tied to them. Of more practical interest, forensic scientists have used them in DNA fingerprinting for many years. However, they remain relatively understudied as sources of genetic variation when compared to other genetic features, such as SNPs, in part because of the computational complexity required to detect them, and partly because they require specialized tools for their analysis. Our lab has developed a useful tool, VNTRseek, to detect VNTRs from high-throughput sequencing data.

With the availability of raw data generated by next-generation, whole-genome sequencing technologies, we can now detect and study features like VNTRs on the genome level. We previously analyzed the genomes of James Watson, a Khoisan individual from South Africa known as KB1, and two trios (one from Utah residents with European ancestry and the other from Yorubans in Nigeria), both sequenced as a part of the 1000 Genomes project. We have now begun work on analyzing more of the available data from various sources, including the 1000 Genomes project and the so-called “Platinum Genomes” provided by Illumina. Here we present the endeavor to build a comprehensive database of VNTRs using these data, the establishment of a robust human genome reference set of tandem repeat loci for detection of VNTRs, an analysis of our results, and the implications that a repository like this may have on the study of complex diseases in humans.

Compiling a Minicircle Genome for Trypanosoma Brucei

Tian Yu[1], Tyler Faits[1], Stefano Monti[2], Ruslan Afasizhev[3], Gary Benson[1]

[1]Bioinformatics Graduate Program, Boston University, Boston, USA
[2]Department of Medicine, School of Medicine, Boston University, Boston, USA
[3]Department of Molecular & Cell Biology, Boston University, Boston, USA

Background:
Trypanosomes, protozoan parasites infecting millions of humans worldwide, have a single, prominent kinetoplast – a large mitochondrion with unique DNA structure. Kinetoplast DNA comprises 10,000 interlinked molecules of circular DNA known as minicircles (1kb) and maxicircles (25kb). The existence of conserved sequence boxes (CSBs) and bend regions in these structures make them poor targets for typical short-read Illumina sequencing to compile a minicircle genome. While suffering from higher error-rate, PacBio sequencing offers near-complete minicircle reads, dramatically reducing the difficulty of sequence assembly.

Description:
We used PacBio to sequence minicircles, and Illumina to sequence the gRNA transcriptome of steady-state Trypanosoma brucei cells. We filtered raw PacBio reads and clustered them based on sequence-similarity and generated consensus sequences based on multiple-alignments for each cluster – validating those sequences by examining gRNA coverage and presence of CSBs. We estimated copy-number for each consensus sequence, and correlated the abundance of each unique minicircle with that of its gRNA transcript. We then used gRNA alignment rates to estimate the completeness of our compiled minicircle genome and created a searchable online database to allow easy access to our results.

Conclusions:
Our analysis yielded 143 minicircles, including 128 unpublished sequences. All identified minicircles contained CSBs and DNA bend regions, which cannot be found in most raw reads. The results showed that the relatively frequent errors of PacBio reads can be corrected through clustering and multiple alignment. The abundant gRNA coverage provided evidence that minicircles are likely to encode multiple gRNAs on both strands.

An Improving SVM-Based Prediction for Dicer Cleavage Site Using Loop/Bulge Length

Yu Bao

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

Background:

Dicer plays an important role in the process of mature miRNA generation, and it is important for the Dicer enzyme to cleave pre-miRNA correctly to generate miRNA with correct seed regions, but the mechanism for Dicer cleavage-site selection is still not fully understood. Until now several researches have been done to solve this problem, among which a recent discovery indicates that the loop/bulge structure plays an important role in the selection of Dicer cleavage sites. Based on this discovery an SVM-based tool named PHDCleav[1] has been established to predict the positions of Dicer cleavage sites, and in this research we present an improved method of PHDCleav which also extends from the loop/bulge structure features.

Method:

In the feature-selection step of PHDCleav, nucleotides in loop/bulge regions were represented differently rom normal nucleotides (A, U, C, G), but PHDCleav failed to consider the effect of the length of the loop/bulge region, which may also play an important role in cleavage-site selection. In this research, we represent nucleotides in loop/bulge regions with different lengths as different patterns from each other as well as normal A/U/C/G. The method of PHDCleav is divided into two parts, nucleotide pattern and binary pattern, while nucleotide pattern counts the nucleotide composition for the sequences of positive and negative class in 3 ways represented as a vector of 5 dimensions (A, U, C, G, L), a vector of 25 dimensions(AA, AC, AU, AG, AL,…LL),a vector of 125 dimensions (AAA,AAU,AAC,AAG,AAL…LLL), and binary pattern represents A as [1,0,0,0], U as [0,1,0,0], C as [0,0,1,0], G as [0,0,0,1], L as [0,0,0,0], where in all cases L represents the nucleotides in loop/bulge regions.
Since the result of PHDCleav showed that the result of binary pattern has a higher accuracy, then we only represent the binary pattern as A[1,0,0,…,0], U[0,1,0,…,0], C[0,0,1,…,0], G[0,0,0,1…,0], L1[0,0,0,0,1…,0]… Ln[0,0,0,0…,1], where L1 to Ln represent the loop/bulge nucleotides with lengths from 1 to n, and all vectors have a length of n+4.

Result:

While the highest prediction result of PHDCleav achieved an accuracy of 86.22% (binary pattern, 5′-arm, shift window with a size of 14 nucleotides), our method has reached a higher accuracy of 88.19%, which indicates that our method outperforms PHDCleav.

[1] PHDcleav: a SVM-based method for predicting human Dicer cleavage sites using sequence and secondary structure of miRNA precursors, Ahmed et al. BMC Bioinformatics 2013, 14(Suppl 14):S9

Bacterial Strain Tracking Across the Human Skin Landscape

Allyson L. Byrd[1,2], Julia Oh[1], Clay Deming[1], Sean Conlan[1], Heidi H. Kong[3], Julia A. Segre[1]

[1]Translational and Functional Genomics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland, USA
[2] Bioinformatics Graduate Program, Boston University, Boston, USA

[3]Dermatology Branch, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, USA

Metagenomics, the genomic sequencing of an entire community of microbiota (bacteria, fungi, viruses), enables an investigation of the full complement of genetic material, including virulence, antibiotic-resistance, and strain-differentiating markers. The high resolution afforded by whole-genome sequencing provides the key to distinguishing between closely related strains – important information since within one bacterial species some strains are beneficial while others are pathogenic to the host.

To differentiate between closely related strains of the same species, I developed a reference-based approach that utilizes both single-nucleotide polymorphisms and genetic content. In this method, I first stringently map sequencing reads against a database of all sequenced strains of an organism, then process the resulting alignment file by a Bayesian statistical framework that reports the closest neighbors of strains present in a sample. Using simulated metagenomic communities, the true positive rate of the pipeline varied between 80 and 90 percent depending on the sequence variability within a species.

To investigate strain-level heterogeneity in metagenomic skin samples from healthy adults, I focused on the two common commensals *Propionibacterium acnes* and *Staphylococcus epidermidis* because of their well-documented sequence variations. Results indicated that an individual's strains of *P. acnes* are shared across multiple sites of the body, and that those strains are more similar within sites of an individual than between individuals. In addition to differences between individuals, *S. epidermidis* also displayed site-specific strains. For example, a single clade of *S. epidermidis* dominated foot samples on all individuals. Overall these results emphasize that both individuality and site-specificity shape microbial communities across the body. Based on longitudinal data, across body sites an individual's strain signatures remain stable for up to a year – a remarkable finding given that the skin is continually washed and exposed to the environment. Future analyses with this resolution should prove particularly powerful in comparing genetic variation of the microbiota between healthy and diseased states.

DemFeature: Upgraded from DemPred

Hao Wang

Institute of Chemistry and Biochemistry, Macromolecular Modeling Group, Free University of Berlin, Germany

In silico models characterizing related molecular compound information with respect to interesting biological problems can effectively accelerate drug development, in particular in an early stage of research for prioritizing compounds before synthesis and clinical tests. It can also help detecting potential side effects early to avoid costly late stage failures. DemFeature is an *in silico* model, which is further developed to DemPred[1], another *in silico* predictive model developed by our group. DemPred was successfully applied building *in silico* models for predicting in vivo half-life and clearance of small-molecule drugs.

Here, we report on the development of DemFeature. The aim is to constitute a specific training subset for a specific test molecule. This method ignores in the training phase molecules of the learning set that are not sufficiently similar or dissimilar to the molecule to be classified. Based on this strategy, ideally, the considered molecule can be more accurately interpreted based on the chemical and biological information of known molecules. In our study, we decided to use the dataset of the Kaggle contest[2] launched by Boehringer Ingelheim in 2012. The Kaggle dataset involves experimental genotoxicity data, which is of high pharmaceutical relevance. The reasons for using this data set to evaluate our model are: (i) It provides a realistic up-to-date prediction scenario for drug classification and (ii) predictions from different professional individuals and groups were submitted that certainly come close to the theoretical limits of what can be achieved for this prediction task.

Our best model so far showed improved statistics, as measured by the Matthews Correlation Coefficient (MCC). The predictive performance of DemFeature is better than of DemPred. Compared with the best submitted model for the Kaggle competition, the predictive performance of DemFeature is the best if applied on the public test set measured by MCC and shows also good prediction performance on private test set.

[1] O. Demir-Kavuk, M. Kamada, T. Akutsu and E.W Knapp: Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features, *BMC Bioinformatics* 12, 412-421 (2011).

[2] Jörg Bentzien, Ingo Muegge, Ben Hamner and David C. Thompson, Crowd computing: using competitive dynamics to develop and refine highly predictive models. *Drug Discovery Today* 18, 472-478 (2013).

Molecular Similarity-Based Predictions of the Tox21 Screening Outcome

Priyanka Banerjee[1,2], Malgorzata N. Drwal[1], Vishal B. Siramshetty[1], Andrean Goede[1], Robert Preissner[1,3], Mathias Dunkel[1]

[1]Structural Bioinformatics Group, Institute for Physiology, Charité – University Medicine Berlin, Berlin, Germany
[2]Graduate School of Computational Systems Biology, Humboldt Univeristy, Berlin, Germany
[3]BB3R – Berlin Brandenburg 3R Graduate School, Free University Berlin, Germany

To assess the toxicity of new chemicals and drugs, regulatory agencies require *in vivo* testing for many toxic endpoints, resulting in millions of animal experiments conducted each year. However, following the Replace, Reduce, Refine (3R) principle, the development and optimization of alternative methods, in particular in silico methods, has been put into focus in the recent years. It is generally acknowledged that the more complex a toxic endpoint, the more difficult it is to model. Therefore, computational toxicology is shifting from modelling general and complex endpoints to the investigation and modelling of pathways of toxicity and the underlying molecular effects. The U.S. Toxicology in the 21st Century (Tox21) initiative has screened a large library of compounds, including approximately 10K environmental chemicals and drugs, for different mechanisms responsible for eliciting toxic effects, and made the results publicly available. Through the Tox21 Data Challenge, the consortium has established a platform for computational toxicologists to develop and validate their predictive models. Here, we present a fast and successful method for the prediction of different outcomes of the nuclear receptor and stress response pathway screening from the Tox21 Data Challenge 2014. The method is based on the combination of molecular similarity calculations and a naïve Bayes machine learning algorithm and has been implemented as a KNIME pipeline. Molecules are represented as binary vectors consisting of a concatenation of common two-dimensional molecular fingerprint types with topological compound properties. The prediction method has been optimized individually for each modelled target and evaluated in a cross-validation as well as with the independent Tox21 validation set. Our results show that the method can achieve good prediction accuracies and rank among the top algorithms submitted to the prediction challenge, indicating its broad applicability in toxicity prediction.

Dynamical Modelling of DNA Damage-Dependent NF-κB Activation

Fabian Konrath, Dorothea Busse, Jana Wolf

Max-Delbrück-Center (MDC) for Molecular Medicine, Berlin, Germany

Environmental factors as well as intracellular processes continuously damage DNA. The degree of DNA damage determines cell fate, DNA repair and survival or in case of irreparable damage senescence or apoptosis to prevent malignant transformation. Since chemotherapy also causes DNA damage, understanding the decision-making process is not only important in terms of tumor formation but also in terms of treatment and occurrence of resistant tumor cells. The decision of whether damaged DNA is repaired or cell death induced is thought to be determined by a complex regulatory network that controls the activity of the oncogene NF-κB and the tumor suppressor p53.

In order to analyze the complex regulatory network we aim to develop a dynamical mathematical model that predicts cell-fate decisions based on the activation status of NF-κB and p53.

As a first step, we defined sub-modules of the overall signaling cascade. To this end, we have developed differential equation models for the modules describing the recognition of DNA double-strand breaks by the sensor proteins PARP-1 and MRN, the subsequent posttranslational modification of the IKK subunit IKKγ in the nucleus as well as the formation of a high-molecular-weight cytosolic complex which is required for an additional modification of IKKγ. The subsequent activation of the IKK complex results in NF-κB activation.

Bayesian Sparse-Group Feature Selection on Gene Expression Data

Edgar Steiger

*Max Planck Institute for Molecular Genetic, Computational Molecular Biology Group,
Berlin, Germany*

Linear regression models in the analysis of gene expression data almost always lead to underdetermined systems because of the large number of genes and a small number of experiments. Methods that impose sparsity on the regression parameters deal with this problem. In the case of grouped variables this sparsity should be imposed both on the group-wise and the within-group level. We propose a Bayesian approach with Spike-and-Slab feature selection in conjunction with categorical distributions to solve these two levels of sparsity. The parameters of the model are matched to the data using the Expectation Propagation algorithm. We applied the method on two different problems: First, on gene expression time-series data, where we revealed causal relationships and delayed effects in gene-regulatory networks, and second, on gene-expression disease classification data, where we predicted outcomes based on pathway-wise grouped genes. Our method was compared to similar approaches and revealed itself to be a fast and dependable alternative. Especially in the case of time-series data it shows greater accuracy in comparisons of synthetic and real data, exhausting the capabilities of the linear model to reconstruct networks while at the same time disclosing the possibly unknown gene-specific delays.

Role of the DPAGT1/β-catenin/YAP Signaling Network in Oral Squamous Cell Carcinoma

Vinay Kartha[1,2], Liye Zhang[2], Samantha Hiemer[3], Maria Kukuruzinska[4], Xaralabos Varelas[3], Stefano Monti[1,2]

[1]Bioinformatics Graduate Program, Boston University, Boston, USA
[2]Division of Computational Biomedicine, Boston University School of Medicine, Boston, USA
[3]Department of Biochemistry, Boston University School of Medicine, Boston, USA
[4]Department of Molecular & Cell Biology, Boston University School of Medicine, Boston, USA

Progression of oral squamous cell carcinoma (OSCC) to metastasis involves complex changes in epithelial cell growth, survival and migration. While the roles of protein N-glycosylation, Wnt/β-catenin and Hippo pathways in cancer have been independently highlighted, the mechanistic interplay between these pathways in promoting tumor metastasis remains less understood. Prior studies have identified this co-dependent homeostatic pathway network to be deregulated in OSCC, playing a vital role in its tumorigenesis. However, identifying exact mediators of these changes still remains a challenging task, crucial to the discovery of novel and lasting OSCC therapeutics. Here, we apply a multi-omic profiling approach to identify potential regulators of OSCC pathogenic pathway activity using a combination of OSCC cell-line gene expression profiles and massive public genomic data. Gene expression signatures pertaining to genetic knockdowns of DPAGT1 - a gene crucial to protein N-glycosylation, and TAZ and YAP – two transcriptional activators involved in the Hippo pathway, were derived using SCC2 cells. Primary human OSCC high-throughput gene expression data from The Cancer Genome Atlas (TCGA) was then projected onto these signatures and analyzed for their association with clinical features including tumor grade and stage. By scoring samples based on their level of pathway deregulation, and additionally leveraging Copy Number Alterations (CNAs) and somatic mutation data, we are able to identify potential upstream genetic regulators of human OSCC development in the context of the DPAGT1/β-catenin/YAP signaling network, paving the way to discovering targets for OSCC therapy.

Accurate Detection of Small Proportion of Cancer through Deep Sequencing Data

Takuya Moriyama

Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan

Deep sequencing is a popular sequencing method used for detection of small fractions of cancer cells. Because sequence error occurs usually with a probability of over 10^{-4}, it becomes a big problem for detection of minimal residual disease (MRD) where cancer cells sometimes exist in proportions of 10^{-4}. Methods that use overlapping pair reads[1] are known to reduce sequence error, but there is no report whether detection of MRD is possible practically, so we evaluated whether or not this method suffices for detection of MRD.

[1] Chen-Harris H1, Borucki MK, Torres C, Slezak TR, Allen JE. Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs. *BMC Genomics*, 14:96, 2013.

Inference of Signal Transduction Pathways from Phosphorylation Data to Identify Targets of Combinatorial Cancer Therapy

Torsten Gross[1,2], Nils Bluethgen[2]

[1]Institute for Theoretical Biology, Humboldt University, Berlin, Germany
[2]Humboldt University, Berlin, Germany

Over-activation of the MAPK/ERK signaling pathway can lead to uncontrolled cell growth and has been associated with many types of cancer. Detailed knowledge about the underlying network has therefore led to the discovery of potent treatments in targeted cancer therapy. However, intrinsic and acquired resistance through network rewiring corrupts their effectiveness, calling for combinatorial therapeutic interventions. The ability to design such appropriate treatments requires the identification and quantification of related kinase interactions. To this end, we conducted an experimental survey of phosphorylation states of selected kinases in different cancer cell lines and measured the response to various types of stimulations and inhibitions. However, these measurements represent the aggregate interplay between all involved kinases and do not directly quantify their direct interactions. To overcome this challenge, we propose a computational method that infers local interaction strengths between pairs of network components from global steady states. Its crucial advantage is a high practicality, as it allows for an efficient and coherent mathematical treatment of large networks, unobserved network components, and noisy data.

Identifying Allele-Specific Expression in Brain Tissue from Parkinson's Disease Patients

Demetrius DiMucci, Rachael Ivison, David Jenkins

Bioinformatics Graduate Program, Boston University, Boston, USA

Allele-specific expression (ASE) occurs when mRNA transcripts from two alleles are expressed at different levels. This disequilibrium results from differential regulation across the two alleles, and such variation in expression abounds for both cis and trans regulatory elements. Parkinson’s disease (PD), a neurodegenerative disease that manifests late in life, affects between seven and ten million people worldwide. In this study, we sought to identify genes undergoing ASE in pre-frontal cortex tissue from PD patients, but not from control patients. Our approach combined SNP microarray and RNA-Seq data from 33 control samples and 11 PD patient samples to identify genes containing at least one heterozygous SNP in any sample. We aligned RNA-Seq reads with GSNAP, a SNP-tolerant aligner, to prevent reference-allele mapping bias. We then used the MBASED software package to identify genes that showed evidence of ASE. To facilitate exploration of complex expression patterns and identification of SNPs experiencing ASE, we created VisualASE, an online resource to search, browse, and visualize heatmaps of allele expression ratios. We found 6,292 genes that displayed significant ASE in at least one individual and 23 genes that showed significant ASE and were significantly overrepresented in patients with PD ($p < 0.05$).

GTEx Assignment of Cancer Cell Lines by Tissue of Origin

Heather Selby[1,4], Zhaleh Safikhani[2], Marieke Kuijjer[4], Nehme El-Hachem[3], Adrian She[2], Rene Quevedo[2,5], Trevor Pugh[2,5], John Quackenbush[4], Benjamin Haibe-Kains[2,5]

[1]Bioinformatics Graduate Program, Boston University, Boston, USA
[2]Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
[3]Institut de recherches cliniques de Montréal, Montreal, Canada
[4]Dana-Farber Cancer Institute, Boston, USA
[5]Department of Medical Biophysics, University of Toronto, Toronto, Canada

Cancer cell lines have transformed our understanding of human cancer cell biology and represent a mainstay of tumor biology and drug discovery. Cell line misidentification, however, voids scientific reproducibility. The cell line MDA-MB-435, for example, was thought to represent metastatic breast cancer, but has the same DNA profile as the M14 melanoma cell line (Yu *et al.* 2015). In our study, we will use the comprehensive Genotype-Tissue Expression (GTEx) dataset to determine the tissues of origin of cancer cell lines. We (Patil *et al.* 2015) and others (Paquet and Hallett 2015; Marchionni *et al.* 2013) developed a novel classification scheme based on gene co-expression analysis that is not only accurate, but also robust to batch effects and inherent biases present in large transcriptomic datasets. Our classification scheme has two major advantages. First, replacing actual gene expression levels with binary, rank-based, pairwise gene expression comparisons in each sample reduces inherent noise and makes them immune to cross-sample variation in technology, platforms, and laboratories. Second, the relative expression reversals on which these gene pairs are scored and chosen as features captures the large differential changes of expression levels in up-regulated and down-regulated genes. Using the healthy tissues in the GTEx data to classify cancer cell lines by their tissue of origin will significantly improve current cancer cell line annotations as well as the reliability of cell-based research, and will benefit biomedical research as a whole. Moreover, our tissue classification scheme will form a framework to identify patient tumors of unknown origin in the clinic.

REFERENCES

Marchionni, Luigi, Bahman Afsari, Donald Geman, and Jeffrey T Leek. 2013. “A Simple and Reproducible Breast Cancer Prognostic Test.” BMC Genomics 14 (1): 336.

Paquet, Eric R, and Michael T Hallett. 2015. “Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype.” Journal of the National Cancer Institute 107 (1): 357.

Patil, Prasad, Pierre-Olivier Bachant-Winner, Benjamin Haibe-Kains, and Jeffrey T Leek. 2015. “Test Set Bias Affects Reproducibility of Gene Signatures.” Bioinformatics , March. doi: 10.1093/bioinformatics/btv157 .

Yu, Mamie, Suresh K Selvaraj, May M Y Liang-Chu, Sahar Aghajani, Matthew Busse, Jean Yuan, Genee Lee, et al. 2015. “A Resource for Cell Line Authentication, Annotation and Quality Control.” Nature 520 (7547): 307–11.

Logical Analysis of Perturbation Data

Katinka Becker[1], Nils Blüthgen[2] and Alexander Bockmayr[1]

[1]Institute for Mathematics, Free University Berlin, Germany
[2]Institute for Pathology, Charité – University Medicine Berlin, Germany

Perturbation experiments are widely used to explore the connectivity of regulatory and signaling networks. The phosphorylation of several proteins is therefore measured under the influence of combinations of stimuli and inhibitors. A main goal of these experiments is to discover differences between mutated and healthy cells, which is an important aspect for drug design. In this talk we will show how signaling networks can be examined by describing logical relations in measured data. Our analysis is built on a perturbation experiment of the EGFR-signaling pathway[1]. Based on a method for logical analysis of data[2] we will investigate the structure of the underlying pathway. We aim to explain the dynamics of the system by discovering patterns of stimuli and inhibitor combinations leading to a particular readout, and to identify differences and similarities between the observations in different cancer cell lines.

[1] Klinger, B., Sieber, A., Fritsche-Guenther, R., Witzel, F., Berry, L., Schumacher, D., Yan, Y., Durek, P., Merchant, M., Schäfer, R., Sers, C. and Blüthgen, N. Network quantification of EGFR signaling unveils potential for targeted combination therapy. *Molecular Systems Biology*, 9: 673, 2013.

[2] Crama, Y., Hammer, P. L., Ibaraki, T. Cause-effect relationships and partially defined Boolean.

Chromatin Accessibility Alterations Due to Plasma Growth Hormone Pulses in Male Mouse Liver

Andy Rampersaud[1], Jeanette Connerney[2], and David J. Waxman[1,2].

[1]Bioinfomatics Graduate Program, Boston University, Boston, USA
[2]Department of Biology, Boston University, Boston, USA

The three-dimensional framework of the genome has revealed the important interface between structure and function with respect to control of genomic regulation. Chromatin accessibility impacts protein-DNA binding events and transcriptional activity, as shown, for example, by DNase-seq analysis that allows for identification of open (accessible) chromatin regions, known as DNase hypersensitive (DHS) sites. Growth hormone (GH) activation of the transcription factor STAT5 induces such changes in chromatin accessibility, and strikingly, sex differences in gene expression characterize more than a thousand genes in mouse liver. GH activation of the transcription factor STAT5 fluctuates in a pulsatile fashion in male liver, but persists relatively uniformly in female liver and plays an essential role in liver sex differences. Here, we investigated whether or not the intermittent pulses of STAT5 activity dynamically alter chromatin structure in male mouse liver. Livers collected from individual male mice were assayed for GH-activated STAT5 activity (classified as STAT5-high vs. STAT5-low) by EMSA analysis. We thus identified 3,004 DHS at STAT5 binding sites discovered by ChIP-seq that are more than two-fold more accessible in STAT5-high livers than in STAT5-low livers (delta-DHS), indicating dynamic opening with each male plasma GH pulse-induced STAT5 binding event. Strikingly, of these 3,004 delta-DHS, 650 were at male-enriched STAT5 binding sites (13.8-fold enrichment, p = 0), and 798 showed male-biased DNase hypersensitivity (12.2-fold enrichment, p = 0) compared to STAT5 status-independent DHSs. These delta-DHSs were depleted at female-biased DNase hypersensitivity sites, indicating a role for the sex-differential STAT5 binding events in establishing and maintaining sex differences in chromatin accessibility. Together, these findings suggest STAT5 binding induces chromatin remodeling leading to DHS opening. *Supported in part by NIH grant DK33765 (to DJW).*

Balanced Benchmarks Show Simple Approaches Succeed with Growing Numbers

Maciej M Kańduła and David P Kreil

Chair of Bioinformatics Research Group, Boku University Vienna, Austria

The majority of data collected in the biomedical sciences more and more come from high-throughput experiments, and data sets are increasingly of genomic scale. The identification and interpretation of biologically relevant patterns in these data, however, remains a bottleneck for both basic and applied research, and has been rate-limiting in the translation of experimental advances to the clinic. A lot of hope is now being placed in the integrated analysis of measurements from different sources, i.e., the joint analysis of different data types. TCGA data sets on cancer are exceptional by collecting systematic matched studies of gene expression and the activities of novel regulators like microRNAs, the accumulation of somatic mutations, the prevalence of DNA methylation, as well as copy number variation, all of which are known to play key roles in this disease. Interestingly, most data is collected for unmatched samples, with more than five times more measurements of cancer tissue data than of normal tissue data.

We are interested in methodological advances in the domain of integrated data analysis. These require careful benchmarking. Studying Kidney Renal Clear Cell Carcinoma as a topical use-case, we have compiled two manually curated lists of annotated pathways. One is constructed to be enriched in true positives, the other is constructed to be enriched in true negatives. Having both positive and negative reference sets is essential for meaningful benchmarks. I here report first results of systematically evaluating individual established methods and a rank-product consensus of a collection of simple algorithms. We can see that analyses of larger unmatched data sets performed better than the carefully matched but smaller data sets. Moreover, even today, a combination of simple tools could outperform the best state-of-the-art algorithm (RTopper).