4TH SCIENTIFIC CBB RETREAT

October 20, 2015

LISTER HILL AUDITORIUM

ORGANIZERS: Michael Galperin, Rezarta Islamaj Dogan, John Spouge

AGENDA

SESSION 1, 10AM - 11:30AM
Evan Bolton "Pubchem, a Chemical Biology Resource"
Chris Lanczycki "VAST+: On To Bigger Things"
Alexey Shaytan "Nucleosomes: From Sequence to Function Via Dynamics"

LUNCH, 11:30AM - 1:00PM

SESSION 2, 1:00PM - 2:30PM
Sergey Shmakov "Discovery of New CRISPR Systems"
Guilhem Faure "Links between mRNA Structure and the Structure of the Encoded Protein"
Eric Nawrocki "Modeling Structural RNA Families with Infernal"

COFFEE BREAK AND POSTER SESSION, 2:30PM - 4:00PM

SESSION 3, 4:00PM - 5:30PM
Anna Panchenko "Perturbation by Sequence Variation: Impacts of Coding Mutations on Proteins and their Interactions"
Amy Sang (Florida State Univ.) "Molecular Signatures and Signaling Pathways of Human Cancers of Breast and Prostate: From Wet Lab to Big Data"
John Wilbur "Words, Probabilities, and Co-Occurrences"

HAPPY HOUR, 6:00PM - 8:00PM
PARVA RESTAURANT & LOUNGE
7904 WOODMONT AVE, BETHESDA, MD 20814

PRESENTERS at the CBB RETREAT

EVAN BOLTON

Pubchem, a Chemical Biology Resource

PubChem (https://pubchem.ncbi.nlm.nih.gov) is an open archive containing community-contributed chemical substance information and their biological activities. Even after eleven years of on-line service, PubChem continues to grow in size [1] and usage [2]. This did not happen entirely by chance. A steady stream of innovations, strategic changes, and community engagement has helped propel PubChem to become a top-ten chemistry website. This talk will give an overview of PubChem and highlight recent changes that are helping to advance the utility of the PubChem resource.

[1] PubChem record counts as of Sep 29, 2015:

Compounds	60,826,515
Substances	157,601,078
BioAssays	1,154,381
Tested Compounds	2,090,729
Tested Substances	3,285,463
RNAi BioAssays	68
BioActivities	229,717,555
Protein Targets	9,953
Gene Targets	57,335

[2] PubChem daily average interactive session count for the week of Sep 20, 2015 from Applog tops 50,000: http://intranet.ncbi.nlm.nih.gov/projects/applog/dashboard/?id=15#from=1442707200&to=1443311999&smoothing=none&tab=0&slice=86400&autoupdate=false

CHRIS LANCZYCKI

VAST+: On To Bigger Things

Since its development in the mid-1990s the Vector Alignment Search Tool (VAST) algorithm has remained a robust, sensitive method to identify structural similarity between proteins. From over 110k sets of 3D coordinate data present in the Protein Data Bank (PDB), VAST has calculated and stored 2 billion structural alignments. In many cases, biological function relies not on individual proteins but an assembly of multiple biopolymers. Recently the PDB began to specify such 'biological assemblies', and assemblies have subsequently become the standard representation of a 3D structure in Structure group public services. As such, a method to generate structural alignments for entire biological assemblies was necessary. The resulting method, called VAST+, analyzes all VAST pairwise structural alignments between proteins from two biological assemblies, looking for a subset that self-consistently aligns pairs of protein chains as they exist in their respective assemblies. Because interesting conformational changes (or similarities) can be localized to small portions of an assembly, current work seeks to focus attention on such regions. VAST+ assembly neighboring has recently replaced VAST as the default public service for finding structural neighbors and can be found at: http://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi.

ALEXEY SHAYTAN

Nucleosomes: from sequence to function via dynamics.

"Nucleosomes are elementary building blocks of chromatin compaction: an octamer of histone proteins wraps about 200 base pairs of DNA into two super-helical turns. Initial conception of nucleosomes as static structures whose sole function is to compact DNA is now giving way to a much more complex understanding of nucleosome as dynamic entities, that actively participate in genome functioning and carry epigenetic markup complementary to the genetic code. The incorporation of alternative histone variants and post-translational modifications into nucleosomes may alter their structure, dynamics and function. In this talk I will survey our work on deciphering the principles of nucleosomes function through the analysis of their dynamics and variability due to different combinations of histone variants.

To address these issues, for the first time we performed long all-atom microsecond molecular dynamics simulations of nucleosomes including linker DNA segments and full-length histones in explicit solvent. We were able to identify and characterize the rearrangements in nucleosomes on a microsecond timescale including the coupling between the conformation of the histone tails and the DNA geometry, as well as behavior of the flexible histone tails. To study the variability in histones we have developed a new version of the histone database (HistoneDB2.0 –with variants) which is publically available on NCBI web site. In addition we constructed structural models of variant nucleosomes (H2A.Z-nucleosomes, centromeric nucleosomes, etc.), analyzed their dynamics, and put forward hypotheses to explain certain cutting edge experimental studies."

SERGEY SHMAKOV

Discovery of new CRISPR systems

Sergey Shmakov, Omar O. Abudayyeh, Kira S. Makarova, Yuri I Wolf, Jonathan S. Gootenberg, Ekaterina Semenova, Leonid Minakhin, Julia Joung, Silvana Konermann, Konstantin Severinov, Feng Zhang, Eugene V. Koonin

Microbial CRISPR-Cas adaptive immunity systems are divided into Class 1, with multisubunit effector complexes, and Class 2, with single protein effectors. Currently, only two Class 2 effectors are known: Cas9, adopted for genome engineering, and Cpf1, recently shown to cleave DNA. We describe two distinct Class 2 CRISPR-Cas systems with effector proteins (C2c1 and C2c3) containing RuvC-like endonuclease domains distantly similar to Cpf1 and a third system, C2c2, with effector containing two predicted HEPN RNase domains. Using RNA-seq, we demonstrate expression and abundant CRISPR RNA (crRNA) production by two C2c1 loci and identify their tracrRNAs. In contrast, two C2c2 loci produce crRNAs independent of tracrRNA. We demonstrate DNA interference by two C2c1 systems expressed in Escherichia coli and determine their PAM sequences. Comparative analysis indicates that Class 2 CRISPR-Cas systems evolved on multiple occasions through combination of Class 1 adaptation modules with effector proteins derived from different mobile elements.

GUILHEM FAURE

Links between mRNA structure and the structure of the encoded protein

Synonymous mutations and mRNA specific structures can modulate the rate of translation, which can substantially affect protein folding. Using the available protein structures from two Eukaryotes and three Prokaryotes, we explored the potential impact of mRNA structure, which was inferred from the mRNA folding energy (dG) profile, on the structure of the encoded protein. We show that dG is positively correlated with several descriptors of protein compactness in Prokaryotes. mRNAs with stable structures that slow down translation typically encode compact proteins. The same relationship was observed in Eukaryotes only after controlling for the GC content, showing that the latter is a suppressor variable. Disordered parts of protein, which are more common in Eukaryotes, are mainly composed of polar residues, which are preferentially encoded by rich GC codons. We further analyzed the relationship between protein structure and mRNA folding energy and found that the size and solvent accessibility of ordered parts of protein positively correlates with the local mRNA folding energy, whereas no such correlation is observed in disordered parts of protein. Thus, there appears to be a robust local relationship between the mRNA folding energy profile and folded portions of a protein. The mRNA structure appears to act as a protein folding controlling device by reducing ribosome speed when the nascent peptide needs time to form and optimize the core structure. This conclusion is compatible with the results of previous studies which indicate that increasing translation speed leads to protein misfolding.

ERIC NAWROCKI

Modeling Structural RNA Families with Infernal

"Non-coding RNAs (ncRNAs) play crucial and diverse roles in a variety of cellular processes including transcription, splicing and gene regulation. Many ncRNAs form stable, evolutionarily conserved three-dimensional structures that are crucial to their function. The Infernal software package implements covariance models (CMs), probabilistic models that take both sequence and structural similarity into account when searching for and creating alignments of homologous RNAs. Infernal is used by the Rfam database to annotate RNAs in genomes and other sequence datasets. The high computational complexity of CM dynamic programming alignment algorithms makes CM methods slow, and has limited their practical application. For example, to make search running times reasonable, Rfam has been forced to use BLAST-based prefilters to prune sequence datasets prior to applying expensive CM methods, even though this sequence-based filtering causes some homologous RNAs that could be recognized using sequence- and structure-based methods to be missed.

I will discuss several acceleration strategies I have implemented in Infernal involving banded dynamic programming and HMM filtering which accelerate CM alignment by 1-3 orders of magnitude and CM homology search by 2-3 orders of magnitude while sacrificing very little sensitivity. These techniques make CM methods fast enough to apply to large sequence datasets, and have eliminated the need for BLAST filters in Rfam searches."

ANNA PANCHENKO

Perturbation by sequence variation: impacts of coding mutations on proteins and their interactions

"Missense mutations can render proteins nonfunctional and may be responsible for many diseases. From the clinical perspective, these non-neutral mutations affecting human health represent the main interest. For some diseases and genes, particularly following the Mendelian inheritance patterns, the causal genotype-phenotype relationship has been already established, while for complex polygenic diseases involving multiple factors it is still unknown.

I will describe new methods and approaches, which we develop to understand the molecular basis of diseases, effects of mutations on proteins and their interactions. I will introduce two new programs/webservers. The first one, MutaBind, allows assessing the impacts of missense mutations on protein-protein interactions and binding affinity. The second method implemented in the server MutaGene describes the background mutagenic processes caused by the intrinsic or extrinsic factors and derives the context-specific empirical models of somatic cancer mutations."

QING-XIANG AMY SANG

Molecular Signatures and Signaling Pathways of Human Cancers of Breast and Prostate: From Wet Lab to Big Data

We discovered a putative cancer biomarker endometase/matrilysin-2/matrix metalloproteinase-26 in human cancers of breast and prostate. We identified a molecular mechanism underlying the aberrant responses to androgen in the androgen-repressed human prostate cancer (ARCaP). Using reversed-phase nano-liquid chromatography coupled to a hybrid linear quadrupole ion trap/Fourier transform ion cyclotron resonance mass spectrometer methods we detected phosphorylated androgen-induced proliferation inhibitor (APRIN). Key identified oncogenic pathways include the mammalian target of rapamycin (mTOR) pathway and the E2F signaling pathway. We also investigated phosphoproteome of human triple-negative breast cancer (TNBC) of a benign breast tissue, a primary breast cancer tissue, and a metastatic breast cancer tissue from the same African American (AA) woman. Notch/Wnt pathway may be a new therapeutic target for TNBC. Moreover, we performed a comprehensive differential gene expression analysis using breast cancer data in the Cancer Genome Atlas (TCGA). In total, 674 RNA transcripts were found differentially expressed between AA and Caucasian American (CA) populations. Resistin, a gene that is linked to obesity, insulin resistance, and breast cancer, was expressed almost five times higher in AA tumors. An uncharacterized, long, non-coding RNA, LOC90784, was down-regulated in AA tumors, and its expression was the lowest in AA TNBC. Network analysis showed increased expression of a majority of components in p53 and BRCA1 subnetworks in AA breast tumor samples, and members of the aurora B and polo-like kinase signaling pathways were also highly expressed. Higher gene expression diversity was observed in more advanced stage breast tumors suggesting increased genomic instability during tumor progression.

W. JOHN WILBUR

Words, probabilities, and co-occurrences

The distributional hypothesis states that words with similar meanings tend to occur in similar contexts (Harris, 1954). Also the idea of a topic is realized by words that co-occur in texts more than expected on a random basis. Even individual words can be tested for their semantic significance to some extent by the how many words they co-occur with. All these ideas are based on the fact that words tend to co-occur at a non-random level. We have developed a way of assigning a probability that two words are related based on co-occurrence and we will describe how this is done, how it can be used for word clustering and how it can be used to search for word ambiguity.

POSTERS at the CBB RETREAT

GELIO ALVES

Mass spectrometry based protein identification with accurate statistical significance assignment

Assigning statistical significance accurately has become increasingly important as metadata of many types, often assembled in hierarchies, are constructed and combined for further biological analyses. Statistical inaccuracy of metadata at any level may propagate to downstream analyses, undermining the validity of scientific conclusions thus drawn.

From the perspective of mass spectrometry based proteomics, even though accurate statistics for peptide identification can now be achieved, accurate protein level statistics remain challenging. We have constructed a protein ID method that combines peptide evidences of a candidate protein based on a rigorous formula derived earlier; in this formula the database P-value of every peptide is weighted, prior to the final combination, according to the number of proteins it maps to. We have also shown that this protein ID method provides accurate protein level E-value, eliminating the need of using empirical post-processing methods for type-I error control. Using a known protein mixture, we find that this protein ID method, when combined with the Soric formula, yields accurate values for the proportion of false discoveries. In terms of retrieval efficacy, the results from our method are comparable with other methods tested.

ROXANNE YAMASHITA

Functional characterization of proteins by domain architecture

Advances in modern sequencing techniques have resulted in an explosion of genomic data. Correctly classifying this new wealth of information can be daunting not only because of the sheer volume of sequence data, but also because the propagation of erroneous and less-than-ideal names and functional characterizations in the current databases gets in the way of functional classification by mere sequence similarity. We are investigating the extent to which protein domain architecture can be utilized to define groups of proteins with similarities in molecular function, and whether we can derive corresponding functional "labels", starting with some of the most common domain architectures found in bacteria. To this end, we have developed an in-house procedure called SPARCLE ("SPecific ARChitecture Labeling Engine") that lets us track and examine specific or sub-family domain architectures, resulting from annotating protein sequences with domain footprints provided by the Conserved Domain Database (CDD), which includes hierarchical classifications for many common domain families. We will discuss how the proteins are grouped into specific architectures, our success rate in assigning functional labels, and the major limitations we have encountered to date. While we will be able to assign functional labels to a large fraction of protein models derived from genome sequences, this effort has the added benefit of pointing out insufficient coverage and resolution of the current protein domain model collections that constitute CDD. We will also discuss alternative procedures that utilize pre-computed domain annotation for clustering protein sequences at a level that is well suited for functional labeling. We hope that this preliminary study will help to identify approaches that facilitate rapid and accurate annotation of genomes with a minimum of manual intervention.

JAIME IRANZO

Virus-host arms race at the joint origin of multicellularity and programmed cell death

Unicellular eukaryotes and most prokaryotes possess distinct mechanisms of programmed cell death. How an "altruistic" trait, such as programmed cell death, could evolve in unicellular organisms? To address this question, we developed a mathematical model of the virus-host co-evolution that involves interaction between immunity, programmed cell death and cellular aggregation. Analysis of the parameter space of this model shows that under high virus load and imperfect immunity, joint evolution of cell aggregation and programmed cell death is the optimal evolutionary strategy. Given the abundance of viruses in diverse habitats and the wide spread of programmed cell death in most organisms, these findings imply that multiple instances of the emergence of multicellularity and its essential attribute, programmed cell death, could have been driven, at least in part, by the virus-host arms race.

CATHERINE FARRELL

From Genes to Functional Elements – Enriching RefSeq Annotation of the Human Genome

The RefSeq database at NCBI (www.ncbi.nlm.nih.gov/refseq/) provides reference sequences for transcripts, proteins and genomes. Its diverse uses include genome annotation, gene identification, variation calling, and functional characterization. To date, the RefSeq group has primarily focused on defining the sequences of genes, pseudogenes and gene products. However, genes occupy only a small fraction of the genome, and the sequences of non-genic functional elements need to be defined. Initiatives such as the ENCODE and Roadmap Epigenomics projects have provided ways to predict the presence of functional elements on the reference genome, but specialized research knowledge and the use of customized genome browsers or tracks are required to assess such elements. Thus, these elements may not be apparent to all users in biomedical research. Therefore, a new RefSeq sub-project will provide reference sequences and genome annotation for experimentally validated functional elements in human and mouse. These elements will include known gene regulatory elements (e.g., enhancers), elements involved in higher-order genome organization (e.g., boundary elements), and elements that are otherwise considered to be of functional importance (e.g., recombination hotspots). Curated records will also be provided for these elements in NCBI's Gene resource (www.ncbi.nlm.nih.gov/gene/) to include summaries, functional attributes, and various metadata from the literature and public databases. This added content will enrich the current scope of our genome annotation, will be particularly useful for GWAS and variant interpretation outside of gene boundaries, will be highly visible and accessible to a wide user base, and is expected to be valuable to research in general.

NARMADA THANKI

Utilizing 3D structure for the annotation of structural motifs in the Conserved Domain Database

The Conserved Domain Database (CDD) is a protein classification and annotation resource comprised of multiple sequence alignments representing ancient conserved domains. CDD protein domain models are curated by NCBI and use 3D protein structure explicitly to define domain extent and the location of conserved core structures, and to provide accurate alignments between diverse family members via structure superposition. CDD also imports external collections such as Pfam and TIGRFAM. Recently, a novel class of annotations labeled as "structural motifs" has been introduced to supplement current capabilities. These annotations define compositionally-biased and/or short repetitive regions in proteins, which are difficult to model as functional domains conserved in molecular evolution. Structural motifs include transmembrane regions, coiled coils, and short repeats with variable copy numbers. For many types of short tandem repeats, a few position-specific score matrices (PSSMs) suffice to annotate more than 90% of the known instances of that structural motif. Unfortunately, a lack of sequence similarity within coiled-coil regions prohibits the development of only a few generic models; therefore, models for coiled-coil regions in the context of specific families have been developed using the Spiricoil Database as a reference.

MARIO A FLORES

Enhancer reprogramming in the mammalian genome

Changes in the transcription binding sites of non-coding regulatory DNA sequences could change their function. Some of the changes will be detrimental for the fitness of the species while other changes might be beneficial or neutral and could become fixed in the population. This "reprogramming" of regulatory elements results in the modification of the gene regulatory landscape during evolution. One type of regulatory DNA sequences are enhancers, which are short DNA sequences that can be bound by transcription factors and activate the transcription of a gene. Enhancers reprogramming has been observed in distinct mammalian genomes and we have identified enhancers that have changed their function during the mammalian evolution. This work allows us to gain understanding of the role of reprogramming in the evolution of the mammalians and possibly vertebrates.

ALEXANDER GONCEARENCO

Probabilistic, context-specific models of somatic cancer mutations

Somatic mutations in cancer cells are caused by intrinsic (e.g. methylation/ deamination) and extrinsic (e.g. DNA binding molecules) mutagenesis factors. Many mutagenic processes are context-specific, where mutational probabilities are determined by the type of mutagenesis and nucleic acid sequence context. While recent findings have uncovered some of these context-dependent mutational patterns, much remains unknown about the process of mutagenesis in cancer and what mutational patterns give cancer cells a selective growth advantage.

By analyzing somatic cancer mutations catalogued in the COSMIC database, we derive context-specific probabilistic models describing the mutagenic process. Missense mutations alter amino acid sequence and may affect cellular fitness and are therefore subject to selection pressure. We assume that synonymous mutations have no effect on fitness and are therefore good approximations of the underlying mutagenic process. The models we derive, along with the tools we have developed to utilize them, are implemented in a dynamic web server, so that they can be accessible to cancer researchers and clinicians.

PHILIPPE YOUKHARIBACHE

Structural plasticity of small beta barrels: Complexity buildup and functional diversification

Small beta barrels form a class of their own. They are among the most ancient proteins and are present in all forms of life. The highly resilient pseudo-symmetric structural framework allows for small structural variations, resulting in an astounding functional plasticity. Small beta barrels are known to interact with proteins, RNA, DNA, and oligosaccharides. Their unique size and shape support various loop lengths and quaternary organizations – from toroid rings of various sizes and compositions that interact with RNA, to membrane channels to fibrils. This on-going work attempts to highlight some fundamental principles of the structure underlying the structural and functional plasticity of small barrels.

SUN KIM

Summarizing topical contents from PubMed documents using a thematic analysis

Improving the search and browsing experience in PubMed is a key component in helping users detect information of interest. In particular, when exploring a novel field, it is important to provide a comprehensive view for a specific subject. One solution for providing this panoramic picture is to find sub-topics from a set of documents. We propose a method that finds sub-topics that we refer to as themes and computes representative titles based on a set of documents in each theme. The method combines a thematic clustering algorithm and the Pool Adjacent Violators algorithm to induce significant themes. Then, for each theme, a title is computed using PubMed document titles and theme-dependent term scores. We tested our system on five disease sets from OMIM and evaluated the results based on normalized point-wise mutual information and MeSH terms. For both performance measures, the proposed approach outperformed LDA. The quality of theme titles were also evaluated by comparing them with manually created titles.

REZARTA ISLAMAJ DOGAN

Identifying genetic interaction evidence passages in biomedical literature

Text describing genetic interactions is difficult to identify due to no simple definition for these interactions and lack of training data. We prepared two manually annotated datasets containing 1793 PubMed® abstract and 1000 full text sentences, respectively. We also built two classification systems to identify genetic interaction evidence, one based on word and context features, and one based on query features used for genetic evidence information retrieval. Both models gave satisfactory results on our manually annotated datasets and we produced four different runs, which were submitted for inclusion in the complete BioC Track system. Identification of genetic interactions in biomedical text is a challenging problem with much work still needing to be done.

BEN BUSBY

NCBI Hackathons: Community Driven User Centered Design of Software to Interface with Genomics Datasets

After we established that genomics professionals from the community could build deployable software for genomic analysis in an educational hackathon setting, (http://dx.doi.org/10.1101/018085) we wanted to go farther with implicitly user-centered software prototyping. Therefore we identified three general categories where available open-source software tools to interface with NCBI datasets were lacking, and set out to build deployable software prototypes, guided by members of the genomics community. The three categories we identified were RNA-seq, translational genomics, and education/democratization of access. We built nine teams of volunteers, three for each category, drawing on the previous experience of these individuals, as well as their motivations for attendance. Drawing from our experiences in the January 2015 hackathon, we understood that the major goal of attendees is to finish deployable software, and therefore we endeavored to enable August attendees to do so by breaking software production into smaller pieces, to enable attendees to work through them in three days and then asking three veteran hackathoners (who are experienced Genomics professionals) to help attendees stitch the work together, and write tests for the resultant software from the other two categories. The resultant software from NCBI hackathons is freely available under a creative commons license at github.com/DCGenomics.

SERGEY SHMAKOV

GUILHEM FAURE

See Presenters section.