Function SIG 2018

Program also available at: https://www.iscb.org/cms_addon/conferences/ismb2018/function.php

Saturday, July 7th

10:15 AM-10:20 AM
Function: Introduction

Room: Columbus IJ

-

10:20 AM-11:00 AM
Using evolutionary information to understand cellular systems

Room: Columbus IJ

- **Kimberly Reynolds**

Presentation Overview: Hide

The analysis of correlated evolution between genes can be used to infer functional interactions between the proteins they encode. Co-evolutionary analyses are often validated by their ability to identify proteins engaged in a physical complex or with a shared metabolic pathway. However, in addition the prediction of interaction, they may also provide valuable information about independence. Using folate metabolism as a case study, we find a pair of enzymes that co-evolve, statistically and experimentally, with one another, but independently from the rest of the pathway. A strategy for identifying groups of proteins that adapt and function as self-contained units would help render cellular systems more tractable and predictable, and suggest practical strategies for metabolic engineering.

11:00 AM-11:06 AM
Continuous evaluation of CAFA

Room: Columbus IJ

- Giuseppe Profiti, University of Bologna and ELIXIR Italy, Italy
- **Castrense Savojardo**, University of Bologna, Italy
- Pier Luigi Martelli, University of Bologna, Italy
- Rita Casadio, University of Bologna, Italy

Presentation Overview: Hide

Critical Assessment of protein Function Annotation algorithms (CAFA) is a scientific challenge ran every two years, consisting in predicting Gene Ontology (GO) terms from protein sequences.

The organizers release a set of protein sequences, participant's predictions should be deposited by the following January, and the evaluation is performed on the experimental annotation accumulated in the following months (at least 6).

A paper with the results is usually published before the following instalment of the challenge: CAFA1 (2010-2011) results were published in 2013, CAFA2 (2013-2014) in 2016, CAFA3 2016-2017 evaluation is still in progress.

Journals like NAR Web Server issue require CAFA results for predictors submitted for publication, however such results are available years after the method was tested in CAFA, and in any case the challenge is run every two years. This leads to a gap: either scientists will use old scores, or they should perform "in house" CAFA-like evaluations.

Given this scenario, we propose to have a centralized continuous evaluation system for CAFA-like assessments. This will help in having consistent and certified scores, clear dataset references and openness. Existing benchmarking platforms like OpenEBench could be exploited in that sense.

11:06 AM-11:13 AM

Commonly under-annotated pathways revealed by structure-based proteome annotation

Room: Columbus IJ

- **Peter Freddolino**, University of Michigan, United States
- Mehdi Rahimpour, University of Michigan, United States
- Chengxin Zhang, University of Michigan, United States
- Yang Zhang, University of Michigan, United States

Presentation Overview: Hide

Computational functional annotation is frequently hampered by the lack of high-identity templates for any new target of interest. We have recently developed a hybrid pipeline combining structural prediction/alignment, sequence alignment, and protein-protein interaction information to obtain combined structure predictions and functional annotations for entire proteomes. We find that our inclusion of structural information makes our workflow unusually strong in performance on difficult targets with limited sequence identity to annotated proteins. Importantly, we also observe that in silico structure prediction can now replace experimental structures for the purposes of functional annotation pipelines. The combined structure/function predictions provided by our pipeline provide an unusual richness of information, and we show several usage cases where insight from these predictions accurately guided follow-up experiments.

Examination of our predictions on several model proteomes reveals a range of commonly over-represented functionalities among poorly annotated proteins, including transcription factors, kinases/phosphatases, and pathogenicity genes. Our

findings provide fundamental new insight into the genetic capacity encoded in proteomes across all domains of life, yield a rich new source of information to seed detailed investigation of the functions of many previously mysterious protein-coding genes, and pave the way for large-scale structure/function annotation for a broader range of proteomes of interest.

11:13 AM-11:20 AM

Large-scale assessment of protein function prediction using heterogeneous ensembles

Room: Columbus IJ

- **Linhua Wang**, Icahn School of Medicine at Mount Sinai, United States
- Jeffrey Law, Virginia Tech, United States
- Shiv Kale, Virginia Tech, United States
- T. M. Murali, Virginia Tech, United States
- Gaurav Pandey, Icahn School of Medicine at Mount Sinai, United States

Presentation Overview: Hide

An effective approach to leveraging the complementarity of methods proposed for protein function prediction (PFP) is to assimilate them into heterogeneous ensembles. We have illustrated that such ensembles can provide significant performance gains over individual PFP predictors. However, our previous work has been limited to a few GO terms due to the computational costs of constructing these ensembles. Here, we report the results of large-scale PFP using heterogeneous ensembles.

Specifically, we constructed and evaluated ensembles for 277 GO terms using 12 diverse base classifiers, and two types of methods, namely stacking with 8 different meta-classifiers and Caruana et al's Ensemble Selection algorithm (CES). Stacking using Logistic Regression (SLR) was the best-performing stacker, and also performed competitively with CES. SLR generally outperformed the best base classifier, with median Fmax improvement increasing with GO term size, namely 0.010 (p=0.21), 0.027 (p=1.1x10-7) and 0.033 (p=1.7x10-10) for small (200-500 proteins), medium (500-1000 proteins) and large (over 1000 proteins) terms respectively. Furthermore, the entire computation took less than 48 hours on a sizeable computing cluster. These results demonstrate that large-scale PFP using heterogeneous ensembles constructed systematically using stacking and CES can be predictive and computationally feasible.

11:20 AM-11:40 AM

Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths

Room: Columbus IJ

- Sergey Nepomnyachiy, Tel Aviv University, Israel
- Nir Ben-Tal, Tel Aviv University, Israel

- **Rachel Kolodny**, University of Haifa, Israel

Presentation Overview: Hide

Reuse – the co-option of segments from unrelated proteins to produce new proteins – underlies protein evolution. Thus, characterizing reuse can offer insights to protein function and evolution. To study reuse patterns, we developed an algorithm that identifies 'themes' – reused segments of similar sequence and structure from protein alignments. Our algorithm finds themes of varying minimal lengths, ranging from 35-200 residues. Using it, we quantify and study reuse in the ECOD database of domains and in the PDB. Indeed, theme reuse is prevalent, and reuse is more extensive when including shorter themes. Structural domains, which are autonomously folded protein parts and the best-characterized form of reuse in proteins, are just one of many, complex and intertwined, evolutionary traces. Others include long themes shared among a few proteins, which encompass and overlap with shorter themes that recur in more proteins. The observed complexity is consistent with evolution by duplication and divergence, suggesting that some of the themes might include descendants of ancestral segments. The observed recursive footprints, where the same amino acid can simultaneously participate in several intertwined themes, has interesting ramifications to characterizing evolution and predicting protein function.

11:40 AM-12:00 PM

Identifying the unknown functions of the minimal bacterial genome

Room: Columbus IJ

- **Magdalena Antczak**, University of Kent, United Kingdom
- Mark Wass, The Univesity of Kent, United Kingdom

Presentation Overview: Hide

Nearly 20 years after the first human genome sequence was published our knowledge and understanding of gene/protein functions remains limited. This is exemplified by the recent identification of the minimal bacterial genome which revealed that one third (149 of 438) of the proteins in this genome were of unknown function. These genes perform essential roles, yet we have no idea of the functions they perform.

We performed an extensive in silico analysis to expand our understanding of the minimal genome. Overall our analysis inferred more informative functions for 59 of the 149 proteins of unknown function. The inferred functions cover multiple areas including protein synthesis, cell division and transport. Our results suggest that >50% of the minimal genome is required for the fundamental life processes of preserving and expressing genetic information. Interestingly we identified many transmembrane proteins in the set of uncharacterised proteins and predict that >70% of these have transporter functions. Our analysis provides insight into the functions of proteins in the minimal bacterial genome, which will now be of interest for experimental characterisation. Further, it highlights the ability to use computational approaches to expand our knowledge and understanding of protein function.

12:00 PM-12:20 PM

Higher-quality metabolic models through improved enzyme annotation algorithms

Room: Columbus IJ

- **Nirvana Nursimulu**, University of Toronto, Canada
- Leon Xu, University of Toronto, Canada
- James Wasmuth, University of Calgary, Canada
- Ivan Krukov, University of Calgary, Canada
- John Parkinson, Hospital for Sick Children, Canada

Presentation Overview: Hide

Metabolic modelling is an effective way to understand factors affecting organisms' growth. Ultimately, such models are key for such purposes as metabolic engineering and drug design. However, sequence similarity searches—typically used to annotate enzymatic function for these models—produce false positive enzyme predictions and fail to consider sequence diversity within enzyme classes. Therefore, various methods have been developed, looking beyond sequence similarity for such elements as domain and catalytic site presence. Here, we start by presenting DETECT (Density Estimation Tool for Enzyme ClassificaTion). In DETECT, the sequence diversity within each enzyme class is captured through density profiles. Then, it calculates likelihood scores for a query sequence given its matches to sequences of different enzyme classes. The use of enzyme-specific score cutoffs calculated from cross-validation gives DETECT higher precision and recall compared to existing methods. It remains that different methods are better suited for predicting certain enzyme classes compared to others. Thus, in a second part, we present an integrative approach for enzyme annotation, where enzyme-specific rules are used for combining the predictions of different tools. Overall, we propose methods for creating high-confidence metabolic models to drive biological discovery.

12:20 PM-12:40 PM

Maize GO Annotation - Methods, Evaluation, and Review (maize-GAMER)

Room: Columbus IJ

- **Kokulapalan Wimalanathan**, Iowa State University, United States
- Iddo Friedberg, Iowa State University, United States
- Carson Andorf, USDA-ARS, United States
- Carolyn Lawrence-Dill, Iowa State University, United States

Presentation Overview: Hide

Maize is both a crop species and a model for genetics and genomics research. Maize GO annotations from Gramene and Phytozome are widely used to derive hypotheses for crop improvement and basic science. The maize-GAMER project is an effort to assess existing maize GO annotations and to improve the quality and quantity of annotations. We designed and implemented a plant-specific reproducible meta-annotator (GO-MAP) that uses diverse component methods including sequence-similarity, domain presence, and three CAFA tools (Argot2, FANN-GO, and Pannzer), to predict GO terms to maize genes and aggregates the predicted annotations as an aggregate dataset. Annotations from Gramene, Phytozome, and maize-GAMER were assessed and compared. Compared to Gramene and Phytozome, the maize-GAMER dataset annotates more genes and assigns more GO terms per gene. The quality of annotations was evaluated using an independent gold-standard dataset (2002 GO annotations for 1,619 genes) from MaizeGDB. In the CC category,

maize-GAMER was the top performer, but it ranked slightly behind Gramene in both MF and BP categories. The maize-GAMER GO annotations have been released publicly, and the containerized GO-MAP tool will soon be released to facilitate annotation of other plant proteomes.

12:40 PM-2:00 PM
Lunch Break

-

2:00 PM-2:20 PM
dbCAN family tools for automated CAZyme (Carbohydrate Active Enzyme) annotation of genomes and metagenomes

Room: Columbus IJ

- **Yanbin Yin**, Northern Illinois University, United States

Presentation Overview: Hide

CAZymes (carbohydrate-active enzymes) are among the most important enzymes for the bioenergy and agricultural industries. CAZyme are also important for human health, because microbes living in the human guts encode the highest percentage of CAZymes to degrade various dietary and host carbohydrates, and changing the dietary carbohydrates will impact the gut microbiota structure and further influence the human health. We have built an online database dbCAN-seq (http://cys.bios.niu.edu/dbCAN_seq) to provide pre-computed CAZyme sequence and annotation data for 5,349 bacterial genomes. Compared to the other CAZyme resources, dbCAN-seq has the following new features: (i) a convenient download page to allow batch download of all the sequence and annotation data; (ii) an annotation page for every CAZyme to provide the most comprehensive annotation data; (iii) a metadata page to organize the bacterial genomes according to species metadata such as disease, habitat, oxygen requirement, temperature, metabolism; (iv) a very fast tool to identify physically linked CAZyme gene clusters (CGCs); and (v) a powerful search function to allow fast and efficient data query. With these unique utilities, dbCAN-seq will become a valuable web resource for CAZyme research, with a focus complementary to dbCAN (automated CAZyme annotation server) and CAZy (CAZyme family classification and reference database).

2:20 PM-2:40 PM
Automated Negative Gene Ontology Based Functional Predictions for Proteins with UniGOPred

Room: Columbus IJ

- **Tunca Dogan**, EMBL-EBI, CanSyL, METU, United Kingdom
- Ahmet Süreyya Rifaioğlu, Middle East Technical University, Turkey
- Rabie Saidi, EMBL-EBI, United Kingdom
- Maria Martin, EMBL-EBI, United Kingdom
- Volkan Atalay, Middle East Technical University, Turkey

- Rengul Atalay, METU, Turkey

Presentation Overview: Hide

Functional annotation of biomolecules in the gene and protein databases is mostly incomplete. This is especially valid for multi-domain proteins. There is a grey area in the protein function data resources, where the truly negative functions and the ones possessed by the protein but have not been discovered or documented yet (i.e. false negatives), reside together. In many cases the information about the functions absent from the target biomolecule can be as important as the assigned functions. It's possible to resolve a portion of this grey area by predicting the functions that the target proteins most probably do not possess. In this study, we present an approach to produce negative functional annotations for protein sequences, along with regular positive associations. Using this approach, we have developed an automated function prediction tool "UniGOPred". The negative prediction performance (recall) was measured as 0.82 for both MF and BP, and 0.66 for CC GO terms (with prediction scores ≤ 0.3), in cross-validation. To the best of our knowledge, the ability of a protein function prediction method to predict negative functions using sequence features is investigated here for the first time. UniGOPred is available as an open access tool at http://cansyl.metu.edu.tr/UniGOPred.html.

2:40 PM-3:00 PM
A Look Back at the Quality of Protein Function Prediction Tools in CAFA

Room: Columbus IJ

- **Morteza Pourreza Shahri**, Montana State University, United States
- Madhusudan Srinivasan, Montana State University, United States
- Upulee Kanewala, Montana State University, United States
- Indika Kahanda, Montana State University, United States

Presentation Overview: Hide

The Critical Assessment of protein Function Annotation algorithms (CAFA) is a large-scale experiment for assessing the computational models for automated function prediction (AFP). The models presented in CAFA have shown excellent promise in terms of prediction accuracy, but quality assurance has been paid relatively less attention. The main challenge associated with conducting systematic testing on AFP software is the lack of a test oracle, which determines passing or failing of a test case; unfortunately, the exact expected outcomes are not well defined for the AFP task. Metamorphic testing (MT) is a technique used to test programs that face the oracle problem by defining metamorphic relations (MRs). An MR determines whether a test has passed or failed by specifying how the output should change according to a specific change made to the input. In this work, we use MT to test five web-based CAFA2 AFP tools by defining a set of MRs that apply input transformations at the protein-level. According to this initial testing, we observe MR violations. Currently, we are working on developing domain-specific MRs based on sequence modifications. In the future, we plan to develop a comprehensive MT tool that is readily available for the AFP community.

3:00 PM-3:20 PM
Updates on CAFA3 and CAFA3.14

Room: Columbus IJ

- **Naihui Zhou**, Iowa State University, United States
- Yuxiang Jiang, Indiana University Bloomington, United States
- Michael Gerten, Iowa State University, United States
- Timothy Bergquist, University of Washington, United States
- Md Nafiz Hamid, Iowa State University, United States
- Deborah A. Hogan, Geisel School of Medicine at Dartmouth, United States
- Kimberley A. Lewis, Geisel School of Medicine at Dartmouth, United States
- Alex W. Crocker, Dartmouth College, United States
- George Georghiou, EMBL-EBI, United Kingdom
- Maria Martin, EMBL-EBI, United Kingdom
- Claire O'Donovan, EMBL-EBI, United Kingdom
- Sandra Orchard, EMBL-EBI, United Kingdom
- Sean D. Mooney, University of Washington, United States
- Casey S. Greene, University of Pennsylvania, United States
- Predrag Radivojac, Indiana University Bloomington, United States
- Iddo Friedberg, Iowa State University, United States

Presentation Overview: Hide

The third CAFA challenge (CAFA3) released its prediction targets in September 2016, and preliminary results were announced in July 2017. CAFA3 featured a term-centric track where predictors were asked to associate a large set of genes (the complete genomes of Candida albicans and Pseudomonas aeruginosa) with a limited set of functions. By collaborating with experimental biologists, we were able to use unpublished whole-genome screen results to evaluate these predictions. To specifically address this question, we hosted an additional challenge CAFA 3.14 (CAFA-Pi) that is dedicated to evaluating term-centric predictions. The final CAFA3 results as well as preliminary CAFA-Pi results will be released and discussed, in addition to highlights of the term-centric evaluations and benchmark proteins.

3:20 PM-3:40 PM

Visualization and annotation of genome-scale metabolic networks

Room: Columbus IJ

- **Ying Zhang**, University of Rhode Island, United States
- Jon Steffensen, University of Rhode Island, United States
- Keith Dufault-Thompson, University of Rhode Island, United States

Presentation Overview: Hide

Metabolism forms the basis for understanding cellular processes in all living organisms and is essential in mediating microbial community and host-microbe associations. Despite the broad application of genome-scale models into studying the function and evolution of metabolic networks, a comprehensive understanding of diverse metabolic processes is still lacking

due to the great complexity and variability of metabolic interactions among different species. To enable the annotation and visualization of complex metabolic networks beyond the scope of existing metabolic pathway databases, we have developed a new algorithm, FindPrimaryPairs, for automatically predicting the element-transferring reactant/product pairs and hence tracing the primary connections of metabolites in metabolic networks. The algorithm has been applied to enable the visualization of metabolic pathways. In the presentation, we will demonstrate new applications of our approach into annotating host-microbe metabolic collaborations and discuss the further integration of protein structural and functional information into studying the evolution of metabolic interactions among different species.

3:40 PM-4:00 PM

Deep Multi-network Embedding for Protein Function Prediction

Room: Columbus IJ

- **Vladimir Gligorijevic**, Flatiron Institute, United States
- Meet Barot, Flatiron Institute, United States
- Da Chen Emily Koo, New York University, United States
- Richard Bonneau, New York University, United States

Presentation Overview: Hide

The prevalence of high-throughput experimental methods has resulted in an abundance of large-scale molecular and functional interaction networks. The connectivity of these networks provide a rich source of information for inferring functional annotations for genes and proteins. An important challenge has been to develop methods for combining these heterogeneous networks to extract useful protein feature representations for function prediction. Most of the existing approaches for network integration use shallow models that cannot capture complex and highly-nonlinear network structures. Thus, we propose deepNF, a network fusion method based on Multimodal Deep Autoencoders to extract high-level features of proteins from multiple heterogeneous interaction networks. We apply deepNF on 6 STRING networks to construct a compact low-dimensional representation containing high-level protein features. We present an extensive performance analysis comparing our method with the state-of-the-art network integration methods such as GeneMANIA and Mashup. In addition to cross-validation, the analysis also includes a temporal holdout validation evaluation similar to the measures in CAFA. Our method outperforms previous methods for both human and yeast STRING networks. Features learned by our method lead to substantial improvements in protein function prediction accuracy, which could enable novel protein function discoveries.

4:00 PM-4:40 PM

Coffee Break

-

4:40 PM-5:00 PM

Proceedings Presentation: HFSP: High speed homology-driven function annotation of proteins

Room: Columbus IJ

- **Yannick Mahlich**, Technical University of Munich, Germany
- Martin Steinegger, Max-Planck-Institute, Republic of Korea
- Burkhard Rost, Technical University of Munich, Germany
- Yana Bromberg, Rutgers University, United States

Presentation Overview: Hide

Motivation: The rapid drop in sequencing costs has produced many more (predicted) protein sequences than can feasibly be functionally annotated with wet-lab experiments. Thus, many computational methods have been developed for this purpose. Most of these methods employ homology-based inference, approximated via sequence alignments, to transfer functional annota-tions between proteins. The increase in the number of available sequences, however, has drasti-cally increased the search space, thus significantly slowing down alignment methods.

Results: Here we describe HFSP, a novel computational method that uses results of a high-speed alignment algorithm, MMseqs2, to infer functional similarity of proteins on the basis of their alignment length and sequence identity. We show that our method is accurate (83% accu-racy) and fast (more than 40-fold speed increase over state-of-the-art). HFSP can help correct at least a 20% error in legacy curations, even for a resource of as high quality as Swiss-Prot. These findings suggest HFSP as an ideal resource for large-scale functional annotation efforts.

5:00 PM-5:06 PM

A New Entropy for Measuring Annotation Consistency with Regards to Protein Signatures

Room: Columbus IJ

- **Rabie Saidi**, EMBL-EBI, United Kingdom
- Maryam Abdollahyan, Queen Mary University of London, United Kingdom
- James Lee, EMBL-EBI, United Kingdom
- Tunca Dogan, EMBL-EBI, CanSyL, METU, United Kingdom
- Ahmet Süreyya Rifaioğlu, Middle East Technical University, Turkey
- Maria Martin, EMBL-EBI, United Kingdom

Presentation Overview: Hide

Both UniProt automatic and manual pipelines use sets of family and domain signatures to infer functional annotations of proteins. Recently, a number of studies have suggested that the same set of signatures does not necessarily imply the same annotations, and that other factors, such as the order of signatures in the protein sequence, may have an impact on its function. However, this impact has not yet been quantified. In this work, we present an information theory based approach to measure the consistency between signature sets and annotations. We propose a new entropy measure which takes the dynamic nature of the annotation process into account by assigning different weights to the presence and absence of an annotation. The results show a high consistency between signature sets and annotations in UniProt Knowledgebase. Apart from quantifying the annotation consistency, our analysis has a few additional implications. One is detection of signatures having complete annotation consistency which can then be used as seeds for generating new annotation rules. Moreover, to

gain a better understanding of the reasons behind inconsistency in some signature sets, we used formal concepts to identify proteins with incomplete annotations and discover potential new subfamilies sharing the same annotations.

5:06 PM-5:13 PM

Predicting the Functions of Actinomyces Universal Stress Proteins

Room: Columbus IJ

- **Taylor Brooks**, Bethune Cookman University, United States
- Remi Jones, Bethune-Cookman University, United States
- Antoinesha Hollman, Jackson State University, United States
- Raphael Isokpehi, Bethune-Cookman University, United States

Presentation Overview: Hide

The bacteria genus Actinomyces are able to grow, reproduce and cause infections in multiple sites of the human body including sites where the conditions for bacteria growth is unfavorable. Genes encoding the universal stress proteins enable bacteria to respond to stress and grow in unfavorable conditions such as limited nutrients and acidic conditions. The goal of the research reported here was to predict the functions of the universal stress proteins encoded in genomes of Actinomyces species. A combination of bioinformatics and visual analytics techniques were used to construct data sets and identify function, transcription direction and operonic arrangement of genes adjacent to the universal stress proteins of Actinomyces. Gene neighborhood analysis revealed a 4-gene operon that includes a USP gene that is associated with the genome of an oral Actinomyces. The operon had function annotation for a sucrose transporter and an enzyme for breakdown of sucrose. The presence of double domain USPs could indicate capacity for biofilm formation. Sugar metabolism is central to the behavior of dental Actinomyces species who are able to persist in biofilms, produce acid and store glycogen-like molecules. Further studies could evaluate the expression levels of the members of the operon in diverse environmental conditions.

5:13 PM-5:20 PM

Identifying protein-protein interaction and protein biochemical cycles based on co-occurrence patterns of orthologous proteins.

Room: Columbus IJ

- **Elad Segev**, Holon Institute of Technology, Israel
- Noam Chapnik, Holon Institute of Technology, Israel
- Roy Yosef, Holon Institute of Technology, Israel
- Edouard Jurkevitch, The Hebrew University of Jerusalem, Israel
- Zohar Pasternak, The Hebrew University of Jerusalem, Israel

Presentation Overview: Hide

99.6% of all known proteins were never tested experimentally or even their expression observed, thus predicting their function relies mainly on comparing their sequence to annotated homologs. However, even with new automated tools for high-throughput functional annotation, the function of many proteins remains unknown since they have no annotated

homologs. In order to identify function and discover protein-protein interaction networks, our study aimed at identifying proteins that are functionally linked to each. We analyzed the co-occurrence patterns of 406,000 orthologous and 118,000 homologous proteins from the fully sequenced non-draft genomes of 4,350 bacteria, 166 eukaryotes and 226 archaea. Validation successfully revealed known networks from various pathways, including nitrogen fixation, glycolysis and ribosome proteins; for example, using the query protein AmoA (a subunit of ammonia monooxygenase), the resulting calculated functional network included AmoB and AmoC, the two other subunits.

This method was found to be both biological and computational practical and efficient , thus, it promises to remain efficient even as more and more genomes are being sequenced.

5:20 PM-5:40 PM

Network-based Gene Function Prediction for Pathogenic Bacteria

Room: Columbus IJ

- **Jeffrey Law**, Virginia Tech, United States
- Shiv Kale, Virginia Tech, United States
- T. M. Murali, Virginia Tech, United States

Presentation Overview: Hide

Thousands of bacterial genomes have been sequenced and annotated. A very large fraction of GO functional annotations for bacterial genes are based on sequence similarity and have not been reviewed by any curator. We sought to examine afresh how well we can predict bacterial gene annotations with experimental evidence using network-based methods.

As a proof of concept, we selected 19 clinically-relevant pathogenic bacteria and created a cross-species network based on protein sequence similarity. We integrated this network with species-specific functional association networks for each pathogen from STRING. We hypothesized that the integrated network would have higher predictive power, despite the large network size and sparsity of annotated nodes.

We evaluated multiple network-based prediction algorithm’s ability to predict experimental annotations, and non-IEA annotations using five-fold cross validation. We found that the SinkSource algorithm consistently outperformed (higher F-max values) GeneMANIA, FunctionalFlow, and other BLAST-based methods. While incorporating STRING with the sequence similarity network did not improve F-max values for non-IEA annotations, the integrated network did yield higher F-max values for experimental annotations (median F-max increased from 0.46 to 0.51 for SinkSource across all BP terms). These results demonstrate that integrating multiple types of data improves predictive power for experimental annotations.

5:40 PM-6:00 PM

Proceedings Presentation: DeepFam: Deep learning based alignment-free method for protein family modeling and prediction

Room: Columbus IJ

- Seokjun Seo, Seoul National University, South Korea
- **Minsik Oh**, Seoul National University, South Korea
- Youngjune Park, Seoul National University, South Korea
- Sun Kim, Seoul National University, South Korea

Presentation Overview: Hide

A large number of newly sequenced proteins are generated by the next-generation sequencing technologies and the biochemical function assignment of the proteins is an important task. However, biological experiments are too expensive to characterize such a large number of protein sequences, thus protein function prediction is primarily done by computational modeling methods, such as profile Hidden Markov Model (pHMM) and k -mer based methods. Nevertheless, existing methods have some limitations; k -mer based methods are not accurate enough to assign protein functions and pHMM is not fast enough to handle large number of protein sequences from numerous genome projects. Therefore, a more accurate and faster protein function prediction method is needed.

In this paper, we introduce DeepFam, an alignment-free method that can extract functional information directly from sequences without the need of multiple sequence alignments. In extensive experiments using the Clusters of Orthologous Groups (COGs) and G protein-coupled receptor (GPCR) dataset, DeepFam achieved better performance in terms of accuracy and runtime for predicting functions of proteins compared to the state-of-the-art methods, both alignment-free and alignment-based methods. Additionally, we showed that DeepFam has a power of capturing conserved regions to model protein families. In fact, DeepFam was able to detect conserved regions documented in the Prosite database while predicting functions of proteins. Our deep learning method will be useful in characterizing functions of the ever increasing protein sequences.

Codes are available at https://bhi-kimlab.github.io/DeepFam.

Monday, July 9th

10:20 AM-10:40 AM

Variation and novelty in evolution: de novo genes arise and enable protein structural innovation

Room: Columbus KL

- **Amir Karger**, Harvard University, United States
- Victor Luria, Harvard University, United States
- Anne O'Donnell-Luria, Broad Institute of MIT and Harvard, United States
- Taran Gujral, Fred Hutchinson Cancer Research Center, United States

- John Cain, Harvard University, United States
- Marc Kirschner, Harvard University, United States

Presentation Overview: Hide

How new protein-coding genes and new protein domains appear in evolution are major questions in biology. While new genes are often built by duplicating existing genes, new genes were recently found to arise de novo from genomic DNA. To understand how new genes may arise de novo, we built a mathematical birth-and-death model based on gene and genome dimensions and dynamic factors such as mutation, recombination and selection. We found most genomes should contain many new genes, with few being maintained. Second, we identified thousands of candidate de novo genes in 20 eukaryotic genomes, using phylostratigraphy and proteomics, and evaluated their predicted biophysical properties. Compared to ancient proteins, new proteins are shorter, more vulnerable to proteases, disordered, likely to bind other proteins, yet less prone to toxic aggregation. To test structural predictions, we performed biophysical experiments comparing human new proteins to ancient proteins. We found that new genes encode short proteins that have distinct structural features and are expressed in brain and male germline, readily providing an avenue for evolutionary testing of function. The continuous creation and destruction of new genes provides a dynamic reservoir of molecular variation that enables genomic exploratory behavior to find new structures and new functions.