Algorithmic Advances and Applications from RECOMB 2017

This Cell Systems Call features invited summaries of 32 of the 38 papers accepted for presentation in the "regular track" at the 2017 Research in Computational Molecular Biology conference. The editors have categorized the summaries into cancer genomics, genetics, mass spectrometry, metagenomics, network analysis, phylogenetics, sequence annotation, sequence informatics, single-cell data analysis, and structural biology.

Cancer Genomics: Discovering Relationships between Cancer Modules

Phuong Dao, Yoo-Ah Kim, Damian Wojtowicz, and Teresa M. Przytycka, NCBI/NLM/NIH; Sanna Madan, University of Maryland; and Roded Sharan, Tel Aviv University

Algorithmic Advance

Analyzing mutual exclusivity and cooccurrences of somatic mutation in cancer and exploring the interplay of such patterns with functional interactions can provide insights into the disease. Considering the relationships of genes within a module and between modules simultaneously, *BeWith* identifies groups of genes with specific mutation patterns within each group while enriched for a distinct pattern between the groups (Dao et al., arXiv: 1704.08889. https://arxiv.org/abs/1704.08889).

Biological Application

We applied *BeWith* in three different settings: BeME-WithFun, which finds a set of modules that are enriched with mutual exclusivity between modules and within which genes are functionally related, identified functionally coherent modules that are relevant for cancer progression and/or are subtype specific; BeME-WithCo, which combines mutual exclusivity between modules and co-occurrence within modules, revealed gene groups that differ with respect to their vulnerability to different mutagenic processes and helped us to uncover pairs of genes with potentially synergistic effects; and BeCo-With-MEFun setting finds cooccurring modules, while genes within modules have both mutual exclusivity and functional interactions.

> …*BeWith* identifies groups of genes with specific mutation patterns within each group while enriched for a distinct relation between the groups.

What's Next?

BeWith is a versatile tool, allowing to investigate various relationships between genes and gene modules. We anticipate the method can be utilized in many applications to identify complex relationships among genes.

Cancer Genomics: Modeling Copy-Number Evolution in Mixed Cancer Samples

Simone Zaccaria, Mohammed El-Kebir, and Benjamin J. Raphael, Princeton University; and Gunnar W. Klau, Heinrich Heine University

Algorithmic Advance

Cancer is an evolutionary process driven by somatic mutations that range in scale from single-nucleotide variants (SNVs) through larger copy-number aberrations (CNAs). To study the evolution of CNAs in cancer genomes, we introduce the copy-number tree mixture deconvolution (CNTMD) problem (Zaccaria et al., In Proc. RECOMB 2017, 318–335. http://dx.doi.org/10.1007/978-3-319-56970-3_20). CNTMD addresses two challenges that distinguish copy-number evolution in cancer from standard phylogenetic-tree reconstruction: first, CNAs typically overlap on the genome, creating complicated dependencies; second, nearly all cancer sequencing is of bulk tissue, meaning that the measurements are a mixture of mutations that are present in different cells. To address these challenges, our integer linear programming algorithm combines an evolutionary model that calculates a phylogenetic tree describing the minimum number of CNAs with a deconvolution algorithm that infers the composition of multiple sequenced samples as different mixtures of subpopulations of cells with unique complements of CNAs.

Biological Application

On simulated data, we outperform existing approaches that focus on either deconvolution or phylogenetic inference. Using multiple samples from a patient with prostate cancer, we infer a phylogenetic tree whose structure has high concordance with a previously published tree that was inferred largely from SNVs. However, our tree provides a higher-resolution reconstruction of copy-number evolution.

> *…specialized phylogenetic techniques…are necessary to study cancer genome evolution.*

What's Next?

The CNTMD algorithm is an example of the specialized phylogenetic techniques that are necessary to study cancer genome evolution. Future work includes more sophisticated evolutionary models and incorporation of single-cell sequencing data.

Cancer Genomics: Network-Based Coverage of Mutational Profiles Reveals Cancer Genes

Borislav Hristov and Mona Singh, Princeton University

Algorithmic Advance

Mutations observed in cancers preferentially target specific pathways and processes, though different genes within them may be mutated across individuals. We introduce a novel framework for discovering cancer genes that identifies small connected components within biological networks—corresponding to functionally related groups of genes—such that large numbers of individuals have at least one of these genes somatically mutated. Our method, nCOP, optimizes an intuitive objective function that balances "coverage" of individuals with the size of the output subnetwork and relies on a single parameter whose value is automatically set after a series of cross-validation tests. We devise an integer linear program to solve the problem optimally and also give a fast heuristic algorithm that works well in practice (Hristov and Singh, Cell Systems *5*, 221–229).

Biological Application

Our approach, nCOP, is more effective in identifying cancer-related genes than previous approaches. Further, by considering per-patient mutational information in the context of networks, our method is able to zoom in on infrequently mutated genes that are nevertheless important for cancer. Our analysis of 24 cancer types reveals several novel genes—mutated across only a handful of individuals—that are promising for further investigation.

> *…by considering per-patient mutational information in the context of networks, our method is able to zoom in on infrequently mutated genes that are nevertheless important for cancer.*

What's Next?

Incorporating information about gene expression and copy-number variation may be especially beneficial in extending the capabilities of our approach.

Cell Systems Call

Cancer Genomics: Unraveling Cancer Genome Structure

Ashok Rajaraman and Jian Ma, Carnegie Mellon University

Algorithmic Advance

Identifying the allele-specific structure of an aneuploid cancer genome is critical to better understand somatic evolution. We present a new network-flow-based algorithm to more accurately separate structural variants (SVs) on different alleles in a given cancer sample. The method builds upon our prior work with Weaver (Li et al., Cell Systems *3*, 21–34) to improve the phasing of SVs, thereby revealing more fine-grained details about the complexity of the underlying genome structure (Rajaraman and Ma, In Proc. RECOMB 2017, 224–240. https://doi.org/10.1007/978-3-319-56970-3_14).

Biological Application

The algorithm was used to identify allele-specific SVs in 36 ovarian cancer genomes from the Cancer Genome Atlas (TCGA) project. We found that, on average, we were able to further phase 60% of previously unphased SVs from Weaver, significantly improving the allele-specific interpretation of complex somatic variants.

…a new network-flow-based algorithm to more accurately separate structural variants (SVs) on different alleles in a given cancer sample.

What's Next?

The algorithm can be further improved by unmixing the subclonal structure of the tumor. Our methods will also allow us to directly approach a major gap in cancer genomic analysis—namely, the interaction of allele-specific SVs and the novel patterns they form. Insights into these interactions may give us a clearer view of these somatic alteration patterns and their unique associations with cancer progression and disease outcome.

Genetics: Auto-learning for Longitudinal Genotype-Phenotype Association

Xiaoqian Wang and Heng Huang, University of Pittsburgh; and Jingwen Yan, Xiaohui Yao, Sungeun Kim, Kwangsik Nho, Shannon L. Risacher, Andrew J. Saykin, and Li Shen, Indiana University

Algorithmic Advance

We proposed a novel machine learning model for genotype-phenotype association studies. Our method automatically extracts interrelationships among longitudinal prediction tasks and uses such auto-learned group structure to enhance endophenotype predictions (Wang et al., In Proc. RECOMB 2017, 287–302. http://dx.doi.org/10.1007/978-3-319-56970-3_18).

Biological Application

We applied our method to the ADNI cohort for Alzheimer's disease study. Compared with five related methods, our model achieves better endophenotype prediction results with respect to root mean square error and correlation coefficient. Moreover, our model selects several important SNPs in the endophenotype prediction of Alzheimer's disease, which has been identified in available literature. The replication of such findings demonstrated the rationality of our prediction results.

Our method automatically extracts interrelationships among longitudinal prediction tasks and uses such auto-learned group structure to enhance endophenotype predictions.

What's Next?

We will analyze the auto-learned group structure among longitudinal endophenotypes in Alzheimer's disease and explore the mutual influence of different regions of interest in brain. Meanwhile, we will look into the causal relationship between important SNPs and identify the most related pathways toward Alzheimer's disease diagnosis. We also plan to apply our model to other diseases, such as cancer.

Genetics: Constructing Accurate Confidence Intervals for Heritability

Regev Schweiger, Eyal Fisher, and Eran Halperin, Tel Aviv University and University of California, Los Angeles

Algorithmic Advance

Estimation of heritability is an important task in genetics. The common way to report the uncertainty in the estimation uses standard errors (SE), which rely on asymptotic properties. However, these assumptions are often violated because of the bounded parameter space, statistical dependencies, and limited sample size, leading to biased estimates and inflated or deflated confidence intervals. FIESTA (fast confidence intervals using stochastic approximation) is a method that builds accurate CIs rapidly, e.g., requiring only several seconds for datasets of tens of thousands of individuals, making FIESTA a very fast solution for all dataset sizes.

Biological Application

We applied FIESTA to construct 95% CIs for heritability of phenotypes using the NFBC dataset (2,520 individuals) and the WTCCC2 dataset (13,950 individuals). The CIs achieve accurate coverage while being orders of magnitude faster than the alternatives.

We applied FIESTA to construct 95% CIs for heritability of phenotypes using the NFBC dataset (2,520 individuals) and the WTCCC2 dataset (13,950 individuals). The CIs achieve accurate coverage while being orders of magnitude faster than the alternatives.

What's Next?

Multiparametric extensions remain a future direction of research: multiple variance components, where the genome is divided into distinct partitions (e.g., according to functional annotations, or by chromosomes), or multiple traits, where several phenotypes are estimated concurrently.

Genetics: Identity-by-Descent Detection at Biobank Scale

Ardalan Naseri and Shaojie Zhang, University of Central Florida; and Xiaoming Liu and Degui Zhi, University of Texas Health Science Center at Houston

Algorithmic Advance

We have developed a new method, RaPID, that is more than 100 times faster than GERMLINE, the fastest identity-by-descent (IBD) detection tool for the past 8 years. The efficiency of RaPID is achieved by a randomization strategy around Richard Durbin's PBWT, which enables this efficient index to work with genomic sequences with errors and mutations. Thanks to PBWT's linear time complexity with the sample size, RaPID is the first computationally feasible method that can infer IBD segments among individuals in a biobank-scale cohort (Naseri et al., bioRxiv, http://dx.doi.org/10.1101/103325).

Biological Application

IBD detection has several applications in association studies and population genetics. We applied RaPID to the data from 1000 Genomes Project and identified both subcontinental and intercontinental IBDs. Most notably, we identified a genetic relationship at 1.5 cM among Finns and Asians, suggesting a possible historical migration event. Overall, our results showed that RaPID is not only fast but also accurate in terms of IBD detection. RaPID will be an essential tool to uncover the rich information in large genotyped biobank data.

RaPID will be an essential tool to uncover the rich information in large genotyped biobank data.

What's Next?

We are working to improve RaPID by tweaking the random projection. We are also applying RaPID to the UK Biobank genotype data of 500,000 participants.

Genetics: Leveraging Fine-Mapping for Summary Statistics Imputation

Yue Wu, Farhad Hormozdiari, and Eleazar Eskin, UCLA; Jong Wha J Joo, Dongguk University-Seoul; and Farhad Hormozdiari, Harvard University

Algorithmic Advance

In this work, we proposed a novel statistical method, CAUSAL-Imp, to impute the genome-wide association studies (GWAS) summary statistics, utilizing the summary statistics of tagged variants and linkage disequilibrium (LD) structure among different variants. CAUSAL-Imp models allelic heterogeneity (i.e., more than one causal variants in a locus) to improve imputation accuracy. Genes that harbor allelic heterogeneity (AH) are widespread in the genome, and modeling AH is essential to detect novel associated variants. In short, CAUSAL-Imp combines the principles of fine-mapping and summary statistics imputation.

Biological Application

We applied CAUSAL-Imp to Northern Finland Birth Cohort (NFBC) dataset and illustrate that CAUSAL-Imp summary statistics are extremely accurate while controlling type I error. In addition, CAUSAL-Imp is applicable to any GWAS datasets because CAUSAL-Imp only requires summary statistics that are widely available.

…CAUSAL-Imp combines the principles of fine-mapping and summary statistics imputation.

What's Next?

We will apply CAUSAL-Imp to the UK Biobank dataset, which has reached a sample size of 500,000 individuals, to impute the summary statistics of untagged variants for detection of novel causal loci. In the algorithmic side, we will extend CAUSAL-Imp to quickly and efficiently perform imputation for loci with thousands of variants.

Mass Spectrometry: When A Single Decoy Is Not Enough

Uri Keich, University of Sydney; and William Stafford Noble, University of Washington

Algorithmic Advance

A key step in the analysis of any tandem mass spectrometry dataset is assignment of statistical confidence estimates. This step is almost always done using a database of "decoy" peptides that are shuffled or reversed versions of real peptide sequences. Keich and Noble (In Proc. RECOMB 2017, 99–116. http://dx.doi.org/10.1007/978-3-319-56970-3_7) propose several algorithmic improvements to the standard decoy-based method for estimating false discovery rates. Their partial calibration method uses a meta-scoring scheme to increase statistical power by increasing the number of decoy databases being searched. Their averaged target-decoy competition method improves upon the standard decoy-based estimation procedure by reducing its liberal bias for small FDR values and decreasing its variability throughout.

Biological Application

These new methods offer more accurate confidence estimates and significantly improved statistical power, allowing proteomics researchers to get much more value out of their data. In one application, analysis of data from *Plasmodium falciparum* yielded 17% more spectra identified at an FDR threshold of 5%.

These new methods offer more accurate confidence estimates and significantly improved statistical power, allowing proteomics researchers to get much more value out of their data.

What's Next?

Searching many decoy databases can be costly, so a key question is how to automatically trade off between running time and accuracy of confidence estimates. It will also be critical to generalize the ideas here to areas such as cross-linking mass spectrometry and analysis of data-independent acquisition data.

Cell Systems Call

Metagenomics: A Bayesian Method for Reconstruction of Viral Populations Characterized by Low Diversity

Soyeon Ahn and Haris Vikalo, The University of Texas at Austin

Algorithmic Advance

We present a novel viral quasispecies reconstruction algorithm, aBayesQR, that employs a maximum-likelihood framework to infer individual sequences in a mixture from high-throughput sequencing data. The search for the most likely quasispecies is conducted on long contigs that our method constructs from the set of short reads via agglomerative hierarchical clustering; operating on contigs rather than short reads enables identification of close strains in a population and provides computational tractability of the Bayesian method.

Biological Application

Performance of the developed method is tested on both synthetic datasets generated by emulating high-throughput sequencing of a viral population and a real HIV-1 dataset. In both settings, aBayesQR outperforms existing techniques in terms of accuracy of the quasispecies size estimation, perfect reconstruction of strains, proportion of correct bases in each reconstructed strain, and the estimation of their abundance; our method particularly stands out when reconstructing a set of closely related viral strains.

…a novel viral quasispecies reconstruction algorithm, aBayesQR…employs a maximum-likelihood framework to infer individual sequences in a mixture from high-throughput sequencing data.

What's Next?

aBayesQR can be extended and applied to the problem of estimating the population size and the degree of variation among the constituent species in related fields such as immunogenetics. On a related note, bacterial populations are characterized by having relatively lower mutation rates than viral and, thus, typically have fewer segregating sites on the sequences in a population. The ability of our method to perform highly accurate reconstruction in such settings should be further investigated.

Metagenomics: A Concurrent Subtractive Assembly Approach for Identification of Disease-Associated Sub-metagenomes

Yuzhen Ye, Wontack Han, and Mingjie Wang, Indiana University

Algorithmic Advance

Recent studies have shown the impacts of microbiomes on human health and diseases. Comparative analysis of metagenomes can be used to detect sub-metagenomes (species or gene sets) that are associated with specific phenotypes (e.g., host status). We developed a concurrent subtractive approach (CoSA) for comparative microbiome studies (Han et al., In Proc. RECOMB 2017, 18–33. http://dx.doi.org/10.1007/978-3-319-56970-3_2). CoSA applies statistical tests to k-mer counts to detect differential k-mers that are likely to be found in differential genes between two groups of metagenomes. Only reads that contain these signature k-mers are extracted and assembled for identification of genes. By focusing on differential reads, CoSA reduces the size of microbiome datasets to be analyzed, allowing faster and better identification of genes of interests.

Biological Application

Application of CoSA to type II diabetes (T2D) microbiome datasets enabled the detection and assembly of genes with consistent abundance difference across samples. An SVM classifier built upon the microbial genes detected by CoSA accurately discriminated patients from healthy controls, and therefore, these microbial genes may serve as potential marker genes for T2D.

Only reads that contain these signature k-mers are extracted and assembled for identification of genes.

What's Next?

We will apply the CoSA approach to detect potential microbial genes for other diseases.

Metagenomics: Fast Metagenomic Binning via Hashing and Bayesian Clustering

Victoria Popic and Volodymyr Kuleshov, Stanford University

Algorithmic Advance

GATTACA is a new lightweight framework for unsupervised binning of metagenomic contigs (Popic, et al., bioRxiv, http://dx.doi.org/10.1101/130997). Similar to other recent approaches, GATTACA clusters contigs based on their coverage profiles across a large cohort of metagenomic samples (using a Bayesian Gaussian mixture model with a Dirichlet prior). Unlike previous methods that rely on read mapping, GATTACA quickly estimates contig coverage features from k-mer counts stored in a compact index built using minimal perfect hash functions. This approach results in up to several orders of magnitude speedup in coverage estimation and metagenomic binning. It also provides a way to index metagenomic samples (e.g., from public repositories like the HMP) offline once and easily reuse them across experiments. The small size of the sample indices (~150 MB per sample) allows them to be easily downloaded and shared.

Biological Application

Researchers can use GATTACA to rapidly identify clusters of contigs belonging to individual microbial species in a metagenome; they can perform this analysis without having to sequence a large cohort of individuals and without having to use a large compute cluster for data analysis.

[GATTACA] results in up to several orders of magnitude speedup in…metagenomic binning.

What's Next?

We envision GATTACA being used to uncover fine species-level structures in metagenomes. This will enable scientists to better assess the diversity of complex bacterial communities that cannot be cultured and to more reliably uncover the presence of pathogenic bacteria.

Network Analysis: Applying Causal Modeling And Bayesian Experimental Design to Inference of Signaling Network Structure

Robert P. Ness, Purdue University; and Olga Vitek, Northeastern University

Algorithmic Advance

We present an algorithm for identifying perturbation experiments that enable causal inference of regulatory pathways using Bayesian networks and Bayesian experimental design. The method encodes pathway knowledge in databases such as KEGG, as well as available datasets on the pathway under study, into a probability distribution on pathway graphs. It identifies a class of pathway structures that cannot be statistically distinguished from one another without perturbations. Then, given a candidate for the true pathway graph, the algorithm identifies a minimal set of perturbations that could distinguish that graph from statistically equivalent graphs. An optimal set of perturbations is found by averaging over the probability distribution on pathway graphs (Ness et al., In Proc. RECOMB 2017, 134–156. http://dx.doi.org/10.1007/978-3-319-56970-3_9).

Biological Application

Prior literature on learning pathway structure from data tends to use well-known pathways as ground truth. This work presents a method for selecting perturbations needed to learning pathway structure when the ground truth is not known, and the experimentalist is interested in novel discovery.

This work presents a method for selecting perturbations needed to learning pathway structure when the ground truth is not known, and the experimentalist is interested in novel discovery.

What's Next?

In ongoing work, we are investigating experimental designs that elucidate reaction rates, going a step further from pathway structure in causal mechanistic description.

Network Analysis: Heterogeneous Information Integration for Drug-Target Interaction Prediction

Yunan Luo, Tsinghua University and UIUC

Algorithmic Advance

Integration of heterogeneous information for predicting drug-target integrations (DTIs) is a promising direction for drug discovery. We introduce DTINet for DTI prediction with the following featured major advances: (1) the generalizability and scalability of integrating multiple types of network data; (2) the effective learning of compact features that best explain the topological properties of nodes in the heterogeneous network; (3) the ability to address the computational challenges that arise from the high-dimensional, incomplete, and noisy biological data (Luo et al., bioRxiv, http://dx.doi.org/10.1101/100305).

Biological Application

DTINet offers a practically useful tool to predict DTIs and may provide new insights into drug discovery and the understanding of mechanisms of drug action. Comprehensive tests demonstrated that DTINet achieved improved prediction accuracy over other existing approaches. We also experimentally validated several novel interactions predicted by DTINet between three drugs and the COX proteins, which suggested new potential applications of these COX inhibitors in preventing inflammatory diseases.

DTINet offers a practically useful tool to predict DTIs and may provide new insights into drug discovery and the understanding of mechanisms of drug action.

What's Next?

DTINet is a scalable framework in that more biological networks of other entities (e.g., gene expression, pathways, and gene ontology annotations) can be easily incorporated into the pipeline. As a versatile method, DTINet can also be applied to various link prediction tasks (e.g., predictions of drug-drug interactions and protein-disease associations).

Network Analysis: Inferring Context-Specific Gene Regulatory Networks by Network Rewiring

Yijie Wang, Dong-Yeon Cho, Hangnoh Lee, Justin Fear, Brian Oliver, and Teresa M. Przytycka, NIH

Algorithmic Advance

We present NetREX (network rewiring using expression), a method to reconstruct gene regulatory network (GRN) given context-specific expression data and a prior GRN that might be largely incomplete or partially incorrect. Thanks to several new algorithmic and conceptual advances, NetREX is the first method to infer GRN that provides a consistent improvement over the prior network. These advances include a novel network optimization approach, overcoming the issue of non-convexity of the optimization problem, and a method that is developed to gauge the relation of the prior GRN to the (unknown) target GRN. We also developed and validated two scoring functions, which are designed for evaluation when a gold standard GRN is not available (Wang et al., arXiv: 1704.05343. https://arxiv.org/abs/1704.05343).

Biological Application

After evaluating NetREX using simulated data and *E. coli* GRN, we utilized it to infer sex-specific GRNs for *Drosophila*. The predicted sex-specific *Drosophila* GRNs were validated using our novel scoring functions and new experimental data on the Doublesex (*DSX*) regulatory interactions.

Thanks to several new algorithmic and conceptual advances, NetREX is the first method to infer GRN that provides a consistent improvement over the prior network.

What's Next?

We plan to use NetREX to compute a regulatory network for the S2 cell—the most frequently used *Drosophila* cell-line. Given the accessibility of context-specific expression data, combining such data with prior knowledge is particularly promising for constructing reliable context-specific GRNs.

Cell Systems Call

Network Analysis: Tracing Genetic 'Partners in Crime' in Cancer Patient Survival Data

Dariusz Matlak and Ewa Szczurek, University of Warsaw

Algorithmic Advance

Epistasis occurs when the contribution of gene alterations to the fitness is non-linear. The type of epistatic interactions with great potential for cancer therapy is synthetic lethality. Inhibitors targeting synthetic lethal partners of genes mutated in tumors can selectively kill tumor, but not normal, cells. Therapy based on synthetic lethality is, however, context dependent, and it is crucial to identify its biomarkers. In our work (Matlak and Szczurek, PLoS Comput. Biol. *13*, e1005626), we used the Lehmann alternatives model to formally connect patient's survival to tumor genotype and its fitness, with the intuition that patients with fit tumors should survive shorter. Based on this model, we devised SurvLRT, a likelihood ratio test for the significance of epistatic interactions. The theory applies to both pairwise and triple interactions. Our approach is the first to use the latter for identification of therapeutic biomarkers.

Biological Application

Inhibitors targeting synthetic lethal partners of genes mutated in tumors are already utilized for efficient and specific treatment in the clinic. SurvLRT can limit the experimental effort of validating epistatic interactions for synthetic lethality-based therapy and for their biomarkers.

SurvLRT can limit the experimental effort of validating epistatic interactions for synthetic lethality-based therapy and for their biomarkers.

What's Next?

Our new project, https://www.mimuw.edu.pl/~szczurek/sl, will mine additional collections of data, coming up with combined evidence of synthetic lethality, boosting predictive power of SurvLRT.

Phylogenetics: Determining the Consistency of Resolved Triplets and Fan Triplets

Jesper Jansson, PolyU; Andrzej Lingas, Lund University; Ramesh Rajaby, NUS; and Wing-Kin Sung, NUS

Algorithmic Advance

A classical algorithm by Aho, Sagiv, Szymanski, and Ullman from 1981 can be used to efficiently merge a set of rooted, binary, distinctly leaf-labeled trees with exactly three leaves each and partially overlapping leaf label sets into a single *supertree* or to determine that no such tree exists due to branching conflicts. We show how the underlying problem's complexity changes when its definition is modified, e.g., by also allowing non-binary trees or "forbidden" trees to be specified in the input. One surprising finding was that things gets *harder* if the output tree's degree is restricted. On the other hand, requiring the input to be *dense* makes the problem much easier.

Biological Application

Inferring a reliable phylogenetic tree from a large dataset is a time-consuming task. The supertree approach (Bininda-Emonds, Trends Ecol. Evol. *19*, 315–322) provides a compromise between accuracy and computational efficiency by divide-and-conquer. Our new results expose the boundary between efficiently solvable and intractable problems within the supertree framework and may therefore be helpful when developing software for constructing large phylogenetic trees.

Our new results expose the boundary between efficiently solvable and intractable problems within the supertree framework.

What's Next?

A few questions remain open. For example, how hard is the following problem variant? Given a set of non-binary phylogenetic trees with three leaves each and overlapping leaf label sets, is there a degree-3 tree that contains all of them?

Phylogenetics: On the Number of Genes Needed in Species Tree Reconstruction

Siavash Mirarab and Shubhanshu Shekhar, University of California, San Diego; and Sebastian Roch, University of Wisconsin - Madison

Algorithmic Advance

Reconstructing species phylogenies with high resolution despite discordance among individual gene trees has motivated researchers to build large *phylogenomic* datasets with thousands of loci. A natural question is how many genes are required for accurate species tree reconstruction. One of the popular methods of species tree reconstruction is called ASTRAL. While the statistical consistency of ASTRAL under a model of genome evolution called the multispecies coalescent was previously known, we have now derived the data requirement of ASTRAL, as well (Shekhar et al., arXiv: 1704.06831. https://arxiv.org/abs/1704.06831). We prove that the number of genes required by ASTRAL grows quadratically with the inverse of the shortest branch length in the species tree, measured in units of coalescent, and also grows logarithmically with the number of species considered. Our simulation analyses show remarkably consistent results with the theory.

Biological Application

The most difficult phylogenetic relationships to resolve are short branches. Thus, the fact that the number of required genes grows polynomially is reassuring. Perhaps, despite large discordance among gene trees, genomes have enough information to resolve even the most difficult relationships.

While the statistical consistency of ASTRAL was previously known, we have now derived the data requirement of ASTRAL, as well.

What's Next?

Will any other method of summarizing gene trees be able to require asymptotically fewer genes than ASTRAL does? The answer remains unknown.

Sequence Annotation: Accurate Predictive Models of the Effects of Non-coding Variants

Jingkang Zhao and Raluca Gordân, Duke University

Algorithmic Advance

Protein-DNA binding models are oftentimes used to predict the effects of non-coding variants on regulatory interactions between transcription factors (TFs) and DNA. When developing such models, the focus is generally on how to represent TF-DNA binding specificity (e.g., using motifs) and how to train the models. But current methods fail to take into account the quality of the DNA-binding data used to train specificity models. We use ordinary-least-squares (OLS) to train regression models of TF-DNA binding from high-throughput *in vitro* data, and we leverage the distributional results associated with OLS estimation to implicitly incorporate data/model quality into our predictions of the effects of non-coding variants on TF binding (Zhao et al., Res Comput Mol Biol. *10229*, 336–352, http://dx.doi.org/10.1007/978-3-319-56970-3_21).

Biological Application

Changes in TF binding due to non-coding variants, as predicted by our OLS models, explained ~50% of the gene expression changes measured in a high-throughput enhancer assay. In addition, we found that pathogenic non-coding variants lead to significant differences in TF binding between alleles compared to common variants.

Changes in TF binding due to non-coding variants, as predicted by our OLS models, explained ~50% of the gene expression changes.

What's Next?

Future work includes adding regularization to our regression models (while maintaining our ability to evaluate the significance of the predicted changes in TF binding) and applying the models to analyze non-coding mutations identified in cancer genomics studies.

Sequence Annotation: A Novel Motif Representation Improves Regulatory Variant Prediction

Yuchun Guo and David K. Gifford, MIT

Algorithmic Advance

KMAC is a *de novo* motif discovery algorithm for computing a novel motif representation, the K-mer Set Memory (KSM), that outperforms existing methods for predicting transcription factor (TF) binding and regulatory variants from DNA sequence. A KSM comprises a set of aligned k-mers that are over-represented at TF binding sites. It captures exact TF-bound sequences without making the positional independence assumption popularized by the position weight matrix (PWM) model. KMAC discovers KSM motifs by density-based clustering and iterative selection and alignment of k-mers (Guo et al., bioRxiv, http://dx.doi.org/10.1101/130815).

Biological Application

KMAC and its computed KSM motifs more accurately predict *in vivo* binding sites than the PWMs and other more complex motif models across a large set of ChIP-seq experiments. In addition, KMAC- and KSM-derived features outperform both PWM and deep learning-model-derived sequence features in predicting differential regulatory activities of expression quantitative trait loci (eQTL) alleles.

…a novel motif representation, the K-mer Set Memory… captures exact TF-bound sequences without making the positional independence assumption popularized by the position weight matrix (PWM) model.

What's Next?

KSMs can be directly incorporated into analyses anywhere PWMs are used. We are developing KSM-based multi-motif learning models for learning the joint binding of TFs to reveal regulatory logic.

Sequence Annotation: Deep Dissection into Translation Dynamics

Hailin Hu, Tsinghua University

Algorithmic Advance

ROSE provides a deep learning-based framework for estimating the likelihood of ribosome stalling on mRNA sequences. Taking advantage of convolutional neural networks with a novel parallel architecture, ROSE outperformed conventional prediction models with a large margin in terms of prediction accuracy (Zhang et al., Cell Systems *5*, 212–220, http://dx.doi.org/10.1016/j.cels.2017.08.004).

Biological Application

ROSE facilitates a genome-wide statistical analysis of putative regulatory factors related to ribosome stalling. Besides confirming several previously established conclusions, we also proposed several novel hypotheses, i.e., a dose-dependent stalling tendency caused by proline residues and a negative correlation between ribosome stalling and codon cooccurrence. Furthermore, genome-wide ribosome stalling landscapes computed by ROSE recovered the functional interplays between ribosome stalling and cotranslational events in protein biogenesis, including protein targeting by the signal recognition particles and protein secondary structure formation. These results established ROSE as a novel computational method to complement the current ribosome profiling techniques and further decipher the complex regulatory mechanisms of translation elongation dynamics encoded in mRNA sequences.

ROSE provides a deep learning-based framework for estimating the likelihood of ribosome stalling on mRNA sequences.

What's Next?

ROSE formulates the problem currently as a binary classification task. We plan to consider a more sophisticated formulation of ribosome stalling prediction next that would enable us to distinguish more specific stalling behaviors, such as duration and causality. We will also apply ROSE to study the mutational effects on translation dynamics and their connections to human diseases.

Cell Systems Call

Sequence Annotation: Inference of the Human Polyadenylation Code

Michael K.K. Leung, Andrew Delong, and Brendan J. Frey, University of Toronto and Deep Genomics

Algorithmic Advance

We predict the usage of polyadenylation sites within a given genomic sequence. Our model architecture is designed to account for competitive aspects of alternative polyadenylation. We experimented with two versions of the model: a neural network with hand-crafted features and a convolutional neural network accepting raw sequences. The convolutional network was consistently more accurate on multiple tasks, suggesting that hand-craft features based on the known biology may not be necessary, as the required signals could be learned directly from the sequence (Leung et al., bioRxiv, http://dx.doi.org/10.1101/130591).

Biological Application

We demonstrate that, without additional training, a model of polyadenylation site usage is effective at other tasks: (1) predicting which polyadenylation site is more likely to be selected in genes with multiple sites, (2) scanning a 3′UTR sequence to find polyadenylation sites, (3) classifying the pathogenicity of variants near polyadenylation sites from the ClinVar database, and (4) anticipating the effect of antisense oligonucleotide experiments on polyadenylation site selection.

> *…hand-craft features based on the known biology may not be necessary.*

What's Next?

Knowing which polyadenylation site is used is not enough to predict impact on protein expression. We plan to model how regulatory elements between alternative polyadenylation sites influence mRNA stability and localization.

Sequence annotation: Ultra-accurate complex disorder prediction

Linh Huynh and Fereydoun Hormozdiari, UC Davis

Algorithmic Advance

In this study, we formalize the problem of building ultra-accurate disorder prediction (UADP) using rare genetic variants (Huynh, bioRxiv, http://dx.doi.org/10.1101/129775). We present a computational framework, Odin (oracle for disorder prediction), for solving this problem in neurodevelopmental disorders. The model and the proposed method are designed to accurately predict the neurodevelopmental disorders for a ''subset'' of affected cases while simultaneously having virtually no false-positive prediction.

Biological Application

Application of our method accurately recovers an additional 8% of autism cases with *de novo* loss-of-function variants in non-recurrently mutated genes with less than 0.5% false-positive prediction. Furthermore, using Odin, we can predict a set of 391 genes that severe variants in these genes can cause autism or other developmental delay disorders.

> *…we formalize the problem of building ultra-accurate disorder prediction (UADP) using rare genetic variants.*

What's Next?

Odin can be extended for predicting the risk of other diseases such as schizophrenia, Alzheimer's, or heart disease. Our work is a step toward the goal of precision medicine and personalized genomics with direct application in clinical settings.

Sequence Informatics: Dynamic Alignment-Free and Reference-Free Read Compression

Guillaume Holley, Roland Wittler, and Jens Stoye, Genome Informatics, International Research Training Group 1906 "DiDy," Faculty of Technology, Center for Biotechnology, Bielefeld University; Faraz Hach;, School of Computing Science, Simon Fraser University, Department of Urologic Sciences, University of British Columbia, Vancouver Prostate Centre

Algorithmic Advance

Large-scale sequencing projects generate an unprecedented volume of genomic sequences, from tens to several thousands of genomes per species. While such sequences are often compressed to reduce storage and transmission costs, no compression tool is currently adapted to consider redundancy and similarity within a collection of genomes instead of a single genome. For this purpose, we developed DARRC (Holley et al., LNCS 10229, 50–65), an alignment-free and reference-free method that compresses sequencing reads dynamically.

Biological Application

DARRC makes use of a novel data structure, the guided de Bruijn graph, to annotate paths in the graph that encode reads. The guided de Bruijn graph enables to update a DARRC compressed archive with reads from similar genomes without full decompression. On a large *P. aeruginosa* dataset totaling 339 Gbp of reads, DARRC provides a 30% compression ratio improvement compared to the best performing state-of-the-art compression method in our experiments.

> *DARRC makes use of…the guided de Bruijn graph to annotate paths of the graph that encode reads…[and] to update a DARRC compressed archive with reads from similar genomes.*

What's Next?

We plan to enhance DARRC with a pattern matching functionality within the compressed data in order to perform large-scale analysis methods, such as variant calling, using multiple compressed genomes.

Sequence Informatics: Enabling Search of Large Sequence Databases with AllSome Sequence Bloom Trees

Chen Sun, Robert S. Harris, and Paul Medvedev, Penn State; Rayan Chikhi, University of Lille

Algorithmic Advance

The Sequence Bloom Tree (SBT) data structure was recently introduced as a way of searching large collections of RNA-seq experiments for transcripts of interest (Solomon and Kingsford, 2016). In Sun et al. (In Proc. RECOMB 2017, 272–286, http://dx.doi.org/10.1007/978-3-319-56970-3_17), we build upon the SBT to introduce the AllSome SBT, which reduces the time to build the index by 53% and to perform a query by up to 85%. Our key insight is that clustering prior to index construction can localize similar experiments into subtrees; such localization can then be exploited to prune the query search space.

Biological Application

The AllSome SBT can be applied to identify publicly available sequencing experiments that contain a certain transcript of interest. Queries could be performed, for example, against all RNA-seq experiments in the NCBI Sequence Read Archive. This can aid biologists in finding papers and datasets of relevance to their particular study. Much like Google search has made vast quantities of information available at a person's fingertips, the AllSome SBT strives to make sequencing databases easily searchable to biologists.

> *Much like Google search has made vast quantities of information available at a person's fingertips, the AllSome SBT strives to make sequencing databases easily searchable to biologists.*

What's Next?

In order to achieve the "googlability" of public sequencing data, many more advances are required. From the algorithmic side, new ideas will be needed in order to make the AllSome SBT scale to the growing size of the SRA.

Sequence informatics: Fast Approximate Long-read Mapping using MinHash

Chirag Jain and Srinivas Aluru, Georgia Institute of Technology; and Alexander Dilthey, Sergey Koren, and Adam M. Phillippy, NIH

Algorithmic Advance

Emerging single-molecule sequencing technologies from Pacific Biosciences and Oxford Nanopore have revived interest in long read mapping algorithms. These sequences can range anywhere from 1 kb to 1 mb in length but sport high error rates in the range of 10%–20%. The high error rate is incompatible with existing mapping algorithms designed for short, highly accurate reads. We describe Mashmap, an algorithm for long-read mapping (Jain et al., In Proc. RECOMB 2017, 66-81. http://dx.doi.org/10.1007/978-3-319-56970-3_5). We demonstrate with Mashmap that a combination of a minimizer index with MinHash identity estimation provides significant benefits in run-time and scalability while achieving precision and recall rates similar to alignment-based methods.

Biological Application

Mashmap allows rapid and accurate mapping of long reads to large reference databases. Empirical results in the paper demonstrate Mashmap's scalability by mapping PacBio metagenomic reads to the entire RefSeq database (838 Gbp) while maintaining high recall values. We hypothesize that such a mapping technique, when combined with nanopore sequencing, could enable real-time genomic analysis of patients, pathogens, cancers, and microbiomes.

> *…provides significant benefits in run-time and scalability while achieving precision and recall similar to alignment-based methods''*

What's Next?

The next step is to generalize this technique for local alignment problems, including split-read and genome-to-genome mappings. This will make Mashmap suitable for much faster analysis of structural variants and homology maps between vertebrae genomes. This is a promising research direction to help address the ever-increasing scale of genomic data.

Sequence Informatics: Fast Sequence Search Using Split Sequence Bloom Trees

Brad Solomon and Carl Kingsford, Carnegie Mellon University

Algorithmic Advance

We present the Split Sequence Bloom Tree (SSBT) for the efficient search of large short-read databases (Solomon and Kingsford, In Proc. RECOMB 2017, 257–271. http://dx.doi.org/10.1007/978-3-319-56970-3_16). We introduce the concept of a "split bloom filter" that minimizes index repetition and prunes more of the search space with each step, leading to faster answers to queries. We further improve this index through the removal of "non-informative" bits, and we use rank and select operations to maintain a consistent hash index across these differentially sized bit vectors.

Biological Application

Using SSBT, it is possible to identify the experiments that are likely to contain the sequences of interest within a database of experiments that otherwise is too large to download in reasonable time. SSBT can index a collection of short-read experiments and search for a sequence ≈5× faster than previous methods while using one-fifth of the space. Since the method is alignment and reference free, it can be used to search for novel sequences. Applications include searching metagenomic experiments, finding structural variants in cancer genomes, and searching RNA-seq collections for expressed genes.

> *SSBT can index a collection of short-read experiments and search for a sequence ≈5× faster than previous methods while using one-fifth of the space.*

What's Next?

We continue to improve the scalability of our methods while also constructing new related methods that are specialized for different biological environments such as cancer tissue.

Cell Systems Call

Sequence Informatics: Joker de Bruijn sequences

Yaron Orenstein and Bonnie Berger, MIT

Algorithmic Advance

We describe a novel approach to design compact sequences that cover all DNA, RNA, or amino acid k-mers using joker characters (Orenstein et al., Cell Systems *5*, 230–236). Joker characters correspond to degenerate nucleotides that are substituted by any nucleotide in the oligo synthesis process. The computational method is based on a greedy heuristic. Its output is improved using an integer linear programming solver. The resulting sequence lengths are very close to the theoretical lower bound.

Biological Application

Shorter sequences that cover all k-mers allow for binding measurements of longer k-mers, as the length of such sequences grows exponentially with k, while the space on the experimental device is limited. Our new design enables binding measurements of DNA 12-mers and amino acid 4-mers, greater than the current state-of-the-art.

We describe a novel approach to design compact sequences that cover all DNA, RNA or amino acid k-mers using joker characters.

What's Next?

Generating universal peptide arrays to cover the complete space of amino acid k-mers, as well as addition of further biological constraints (e.g., covering a k-mer or its reverse k-mer as motivated by peptide synthesis) into array design considerations.

Sequence Informatics: Speeding up FM indices with EPR-Dictionaries

Christopher Pockrandt, International Max Planck Research School of Computational Biology and Scientific Computation; and Marcel Ehrhardt and Knut Reinert, Free University of Berlin

Algorithmic Advance

EPR-Dictionaries (Pockrandt et al., In Proc. RECOMB 2017, 190–206, https://link.springer.com/chapter/10.1007/978-3-319-56970-3_12) replace wavelet trees in FM indices and allow searches in optimal running time by eliminating the logarithmic factor in the alphabet size. The practical experiments correlate with the theoretical improvement, which leads to speedup factors for bidirectional indices of 2× for DNA and 4.6× for protein alphabets compared to wavelet trees. The space consumption grows only by a factor of o(sigma); hence, the dominating term equals the one from wavelet trees.

Biological Application

Fast bidirectional indices speed up many bioinformatic applications that can benefit from switching search directions such as for approximate string searching used by read mappers or for identifying lncRNA based on their secondary structure.

…speedup factors for bidirectional indices of 2× for DNA and 4.6× for protein alphabets…

What's Next?

With bidirectional indices becoming faster and faster, we can push the bounds of approximate string searching. Backtracking approaches in indices are only feasible to up to two to three errors due to the exponential search space. By continuing the work of Kucherov et al. (In Proc. CPM 2014, 222–231, http://dx.doi.org/10.1007/978-3-319-07566-2_23) and the use of EPR-dictionaries, we can make string searching faster by a magnitude and push the bound of feasible number of errors even further.

Single-Cell Data Analysis: *E Pluribus Unum*, United States of Single Cells

Joshua D. Welch and Jan F. Prins, University of North Carolina

Algorithmic Advance

We developed MATCHER, an approach for matching single-cell profiles across different measurement types (Welch et al., Genome Biol. *18*, http://dx.doi.org/10.1186/s13059-017-1269-0). MATCHER uses manifold alignment to learn a low-dimensional representation in which different modalities are directly comparable. Because MATCHER learns a generative model, it can infer single-cell multiomic profiles from measurements performed on different single cells.

Biological Application

MATCHER provides a way to "unite the states" revealed by many different transcriptome and epigenome measurements into a single picture of cell heterogeneity. We inferred multiomic profiles using single-cell gene expression, chromatin accessibility, histone modification, and DNA methylation data from embryonic stem cells and induced pluripotent stem cells. These profiles revealed several insights about the dynamic interplay among epigenetic and transcriptional states during biological processes. For example, we identified some transcription factors whose activity was regulated by both chromatin accessibility changes and transcriptional downregulation, in contrast to others that were regulated primarily at the chromatin level. We also began to unravel the relative ordering of transcriptome and epigenome changes, finding that gene expression changes temporally precede DNA methylation changes during iPSC reprogramming.

MATCHER provides a way to "unite the states" revealed by transcriptome and epigenome measurements into a single picture of cell heterogeneity.

What's Next?

We plan to explore additional types of single-cell data, including chromatin interaction and proteomic measurements. Additionally, we are investigating how to extend MATCHER to align cell profiles from branching biological processes.

Structural Biology: Artificial Intelligence Solves Membrane Protein Structures

Jinbo Xu, Toyota Technological Institute at Chicago

Algorithmic Advance

Wang et al. (Cell Systems *5*, 202–211) describes an efficient deep transfer learning (DL) method for membrane protein (MP) contact prediction by learning MP contact patterns from thousands of soluble proteins, overcoming the challenge of insufficient solved MP structures for model parameter estimation. The DL algorithm formulates contact prediction as an image pixel-level labeling problem and integrates two deep residual networks to predict contacts from input features.

Biological Application

The DL method works well on MP contact prediction, outperforming pure coevolution methods (e.g., CCMpred and Evfold) by a large margin and allowing accurate contact prediction for small-sized protein families. The improved contact prediction can correctly fold 280 of 510 non-redundant MPs and generate 3D models for 57 and 108 MPs with RMSD less than 4Å and 5Å, respectively. In the blind CAMEO test (http://www.cameo3d.org), this DL method predicted a 3D model with RMSD ~2Å for a test target (PDB: 5h35E) of 212 residues. It is estimated that the DL method can correctly fold 1,345–1,871 reviewed human multi-pass MPs, which shall facilitate experimental structure determination of MPs and discovery of drugs targeting at MPs.

> *…an efficient deep transfer learning (DL) method for membrane protein (MP) contact prediction by learning… from…soluble proteins…*

What's Next?

Further development includes integrating more MP-specific information and extending the DL method to predict interresidue distance instead of contacts for higher-resolution 3D modeling.

Structural Biology: Enabling Provable Protein Design over Large Sequence Spaces

Adegoke A. Ojewole, Jonathan D. Jou, Vance G. Fowler, and Bruce R. Donald, Duke University

Algorithmic Advance

Protein design algorithms that optimize for binding affinity are improved when they incorporate ensemble-based design, continuous conformational flexibility, and provable guarantees. However, all previous methods with these desired properties search over an exponential number of sequences one at a time. We introduce a new protein design algorithm, *BBK** (Ojewole et al., In Proc. RECOMB 2017, 157–172, http://dx.doi.org/10.1007/978-3-319-56970-3_10), that retains all three design principles yet efficiently computes the tightest-binding sequences. *BBK** efficiently and provably bounds the binding affinities of a combinatorial number of sequences without computing the binding affinity for any one sequence. Ultimately, *BBK** enables protein designs that span more mutable residues, model more side-chain and backbone flexibility, and search a significantly larger space of possible sequences.

Biological Application

*BBK** is a key improvement upon the single-sequence *K** algorithm (Georgiev et al., J. Comput. Chem. *9*, 1527–1542; Lilien et al., J. Comput. Biol. *12*, 740–761), which has been successfully applied to design proteins in several areas, including broadly neutralizing HIV antibodies. We used *BBK** to redesign the *S. aureus* FNBPA-5:fibronectin interface, and we showed that a flexible backbone model favors binding in different sequences than the fixed backbone model does.

> *BBK** enables protein designs that span more mutable residues, model more side-chain and backbone flexibility, and search a significantly larger space of possible sequences.

What's Next?

*BBK** is readily extended to enhance design with sparse residue interaction graphs, as well as designs with additional backbone flexibility. Furthermore, *BBK** can prospectively predict resistance mutations to drugs in many drug targets, including bacteria, viruses, and cancer.

The following papers were accepted at RECOMB 2017, but summaries are not included here.

A Flow Procedure for the Linearization of Genome Variation Graphs

David Haussler, Maciej Smuga-Otto, Benedict Paten, Adam Novak, Sergei Nikitin, Maria Zueva, and Miagkov Dmitrii

Resolving Multicopy Duplications de Novo Using Polyploid Phasing

Sudipto Mukherjee, Mark Chaisson, Sreeram Kannan, and Evan Eichler

Superbubbles, Ultrabubbles, and Cacti

Benedict Paten, Adam Novak, Erik Garrison, Eric Dawson, and Glenn Hickey

Boosting Alignment Accuracy by Adaptive Local Realignment

Dan DeBlasio and John Kececioglu

Reconstructing Antibody Repertoires from Error-Prone Immunosequencing Datasets

Alexander Shlemov, Sergey Bankevich, Andrey Bzikadze, Yana Safonova, and Pavel Pevzner

A Bayesian Framework for Estimating Cell Type Composition from DNA Methylation Without the Need for Methylation Reference

Elior Rahmani, Regev Schweiger, Liat Shenhav, Eleazar Eskin, and Eran Halperin