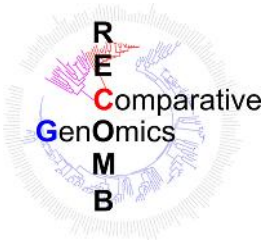

RECOMB-CG

19-22 October 2014

New York – Cold Spring Harbor

IBM **Research**

CSH Cold Spring Harbor Laboratory

PHILIPS

SIMONS FOUNDATION

Posters

1 Ogun Adebali, Davi Ortega and Igor Zhulin. *CDvisto: a Comprehensive Domain Visualization Tool*

2 Charlotte Darby, Maureen Stolzer and Dannie Durand. *What's in a name? An expanded classification of xenologs*

3 Minli Xu, Jeffrey Lawrence and Dannie Durand. *Comparative genomics sheds light on the evolution and function of the Highly Iterative Palindrome -1 motif in Cyanobacteria*

4 Manuel Lafond, Emmanuel Noutahi, Jonathan Séguin, Magali Semeria, Nadia El-Mabrouk, Laurent Gueguen and Eric Tannier. *Gene Tree Correction with TreeSolver*

5 Anna Paola Carrieri and Laxmi Parida. *SimRA: Rapid & Accurate Simulation of Populations based on Random-Graph Models of ARG*

6 Francesco Abate, Sakellarios Zairis, Elisa Ficarra, Andrea Acquaviva, Chris Wiggins, Veronique Frattini, Anna Lasorella, Antonio Iavarone, Giorgio Inghirami and Raul Rabadan*. Pegasus: annotation and prediction of oncogenic gene fusion events as a supervised learning task*

7 Daniel Doerr, Jens Stoye and Katharina Jahn. *Discovering common intervals in multiple indeterminate strings*

8 Guillaume Holley, Roland Wittler and Jens Stoye. *Bloom Filter Trie - a data structure for pan-genome storage*

9 Manfred Klaas, Paul Cormican, Thibauld Michel and Susanne Barth. *Genotyping by sequencing of a collection of Miscanthus spp. accessions*

10 Han Lai and Dannie Durand. How much are you willing to pay? Selecting costs for reconciliation with duplication and transfers

11 Siavash Mirarab, Rezwana Reaz, Md. Shamsuzzoha Bayzid, Théo Zimmermann, M. Shel Swenson and Tandy Warnow. ASTRAL: fast and accurate species tree estimation from gene trees

12 Siavash Mirarab, Nam-Phuong Nguyen and Tandy Warnow. PASTA: ultra-large multiple sequence alignment

13 Alexandra Dana and Tamir Tuller. The effect of tRNA levels on decoding times of mRNA codons

14 Ghada Badr and Arwa Alturki. CompPSA: A Component-Based Pairwise RNA Secondary Structure Alignment Algorithm

15 Robert Aboukhalil, Joan Alexander, Jude Kendall, Michael Wigler and Gurinder Atwal. Single-cell sequencing: How many is many enough?

16 Cedric Chauve, Yann Ponty and João Paulo Pereira Zanetti. Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach

17 Sapna Sharma and Klaus F. X. Mayer. *Genome and sequence characteristics indicate frequent introgressive hybridization events in monocots and dicots*

18 Nina Luhmann, Cedric Chauve, Jens Stoye and Roland Wittler. *Scaffolding of Ancient Contigs and Ancestral Reconstruction in a Phylogenetic Framework*

19 Ghada Badr and Haifa Alaqel. *Genome Rearrangement for RNA Secondary Structure Using a Component-Based Representation: An Initial Framework*

20 Di Huang and Ivan Ovcharenko*. Identifying risk-associated regulatory SNPs in ChIP-seq enhancers*

21 Kevin Emmett and Raul Rabadan. *Characterizing Horizontal Gene Transfer in Microbial Evolution using Topological Data Analysis*

22 Mehmet Gunduz, Esra Gunduz, Omer Faruk Hatipoglu, Gokhan Nas, Elif Nihan Cetin, Bunyamin Isik and Ramazan Yigitoglu. *Role of p33ING1b in Head and Neck Cancer*

23 Pedro Feijao, Fábio V Martinez, Marília Braga and Jens Stoye. *The Family-Free Double Cut and Join and its application to ortholog detection*

24 Corey Hudson and Kelly Williams. *LearnedPhyloblocks: Novel Genomic Islands through Phylogenetic Profiling*

25 Philip Davidson, Luisa Hiller, Michael T. Laub and Dannie Durand. *Tracking the Evolution of a Signal Transduction Pathway Architecture with Comparative Genomics*

26 Krister Swenson and Mathieu Blanchette*. Linking Genome Rearrangements and Chromatin Conformation*

27 Filippo Utro, Deniz Yorukoglu, David Kuhn, Saugata Basu and Laxmi Parida. *Topological Data Analysis to detect population admixture in recombining chromosomes*

28 Yee Him Cheung, Nevenka Dimitrova and Wim Verhaegh. *Achieving Cross-Platform Compatibility of Gene Expression Data*

29 Filippo Utro, Daniel E. Platt and Laxmi Parida. *K-mer Analysis of Ebola sequences differentiates outbreaks*

Poster 1

CDvist: a Comprehensive Domain Visualization Tool

Ogun Adebali[1,2 §], Davi R. Ortega[1,2] , Igor B. Zhulin[1,2]

[1] Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37861, USA, [2] Department of Microbiology, University of Tennessee, Knoxville TN 37996, USA

[§] Corresponding author and poster presenter: oadebali@vols.utk.edu

Abstract

The study of a novel protein starts by obtaining information about the protein using computational methods. The public databases of protein sequences provide a framework for theoretical predictions of function and structure of biomolecules. Based on this wealth of information, specialized datasets of protein domain models are maintained to facilitate protein domain recognition in newly sequenced proteins. Several publically available webservers, HMMER, CD-search and HHpred utilize a variety of algorithms to predict protein domains in sequences based on similarity searches against these datasets. Despite the power and the popularity of these algorithms, none of the available services combines batch querying, consistent visualization scheme and a comprehensive retrieval of protein domain information, especially for multi-domain proteins. More specifically, all these services operate on a whole protein sequence given as input, which may bias results towards more conserved domains and may leave significant protein regions without a match to a known protein domain profile. We propose that domain coverage in multi-domain proteins can be dramatically increased by automated exhaustive search of protein regions without significant match against a variety of databases. We have developed CDvist (Comprehensive Domain Visualization Tool), which combines the power of existing algorithms (HMMER, RPS-BLAST, HHsearch, HHblits) and protein domain databases to a user-friendly visualization framework. To increase domain coverage, rather than using the entire sequence, CDvist iteratively identify regions without significant domain match and submits each of these segments to similarity search against a pre-determined sequence of databases until the entire protein sequence is covered or all databases have been searched. Our web-server allows bulk querying at a high speed enabled by a parallel processing environment. A custom JavaScript module is implemented to represent results in a comprehensive, biologist-friendly manner. We designed CDvist web-server to be used primarily by experimentalists, who are interested to learn more about protein or protein sets of choice. However, it is also attractive to computational biologists due to its bulk querying and JSON formatted export features.

Poster 2

What's in a name? An expanded classification of xenologs

Charlotte Darby [1*], Maureen Stolzer [1*], Dannie Durand [1§]

[1] Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[*] Equal author contribution

[§] Corresponding author durand@cs.cmu.edu

CD is the poster presenter cdarby@andrew.cmu.edu

Abstract

Horizontal gene transfer occurs when a species acquires a gene from a source other than its ancestor. This phenomenon is a fundamental process of gene family evolution in prokaryotes, and evidence is mounting that it occurs in eukaryotes. Growing literature describes the diversity and complexity of gene family histories involving horizontal gene transfer. However, the nomenclature currently available to describe homology relationships when transfer is implicated remains ambiguous. Careful classification of horizontally transferred genes is essential for gaining insight into complex evolutionary processes. Precise characterization is also important because gene homology is frequently used to predict gene function.

Gray and Fitch (Mol. Biol. Evol. 1983) coined the term "xenolog" to describe "clearly homologous" relationships involving genes of foreign origin. In his landmark review, Fitch (Trends Genet. 2000) defined xenology as "the relationship of any two homologous characters whose history, since their common ancestor, involves an interspecies (horizontal) transfer of the genetic material." Current terminology based on this definition would label all genes related through a transfer event as xenologs, not distinguishing among the different homologous relationships involving transfer that can occur.

Expanding upon Fitch's definition, we propose a classification scheme that offers much-needed precision for describing xenologous relationships. Our scheme distinguishes between gene pairs related by transfer alone and genes related by both duplication and transfer. Additionally, our system accounts for the inherent asymmetry of horizontal transfer by differentiating between the donor and recipient species. Whether or not both genes are in the same species is also taken into account. It further considers when genes diverged relative to the divergence of the species in which we observe those genes. Further, we define formal rules that unambiguously assign gene pairs to the xenolog subtypes in our classification. These rules are based on gene tree-species tree reconciliation and have been implemented in prototype software. To show the importance of these distinctions, we apply our conceptual framework to a representative published example: the S. cerevisiae biotin synthesis pathway. This example, one of many that can be found in the literature, demonstrates how our terminology facilitates interpretation of functional relationships between xenologs.

Poster 3

Comparative genomics sheds light on the evolution and function of the Highly Iterative Palindrome -1 motif in Cyanobacteria

Minli Xu [1], Jeffrey G. Lawrence [2], Dannie Durand [3§]

[1] Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213 [2] Department of Biological Science, University of Pittsburgh, Pittsburgh, PA 15213 [3] Department of Biological Science, Carnegie Mellon University, Pittsburgh, PA 15213

[§] Corresponding author durand@cs.cmu.edu

MX is the poster presenter minlix@andrew.cmu.edu

Abstract

The Highly Iterative Palindrome-1 (HIP1), an octamer palindromic motif (GCGATCGC), is highly abundant in a wide range of cyanobacterial genomes from various habitats. HIP1 frequency can be as high as one occurrence per 350 nucleotides, which is rather astonishing considering that at this frequency, on average, every gene in that genome will be associated with more than one HIP1 motif. HIP1 was first identified in the early 1990s, yet its functional and molecular roles are still not understood. No mechanism or biological system has been identified that explains this level of prevalence. More discouraging, it is still not clear whether HIP1 has a function, or whether HIP1 abundance is an artifact of some neutral process, such as DNA repair or transposition.

Here we present results from genome scale analyses that provide the first evidence that HIP1 motifs are under selection. We estimate the expected HIP1motif frequency, taking into account the background tri-nucleotide frequency in the genome, and showed that observed HIP1 frequencies are as much as 100 times higher than expected. This HIP1 motif enrichment is observed in both coding and non-coding regions. Analyses of alignments of genomes with Ks values ranging from 0.02 to 0.59 further showed HIP1 motif conservation in homologous sequences. The level of HIP1 conservation is significantly higher than the conservation of control motifs, i.e., other octamer palindromes with the same GC content. To show that such conservation is not merely a result of codon usage, we demonstrated that codons in HIP1 motifs are more conserved than the same codons found outside HIP1 motifs. Our results, taken together, are consistent with selection acting on HIP1 motifs. We provide the first concrete evidence for the hypothesis that the abundance of HIP1 motifs is related to biological functions, rather than to some neutral process.

Poster 4

Gene Tree Correction with TreeSolver

Manuel Lafond[1*], Emmanuel Noutahi[1], Jonathan Séguin[1], Magali Semeria[2], Nadia El-Mabrouk[1*], Laurent Gueguen[2] and Eric Tannier[2,3]

[1] Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, QC, Canada H3C 3J7 [2] Laboratoire de biométrie et biologie évolutive, UMR CNRS 5558, Université Lyon I, F-69622 Villeurbanne, France [3] INRIA Grenoble Rhône-Alpes, F-38334 Montbonnot, France

[§] Corresponding authors and poster presenters

ML: lafonman@iro.umontreal.ca

N E-M: mabrouk@iro.umontreal.ca

Abstract

We present TreeSolver, a new integrated framework for gene tree correction accounting for local mutations at the sequence level, as well as global mutations affecting gene content and order. In the same vein of recently developed software such as TreeFix, a neighborhood of an input tree is explored, and a correction selected on genome-level criteria is accepted only if it is statistically equivalent to the original tree. However, while a tree neighborhood is explored by previous algorithms in a stochastic way, we take a deterministic and more targeted approach by focusing on the problematic parts of the tree: weakly supported edges and nodes.

Poster 5

SimRA: Rapid & Accurate Simulation of Populations based on Random-Graph Models of ARG

Anna Paola Carrieri[1], Laxmi Parida[2§]

[1] Università degli studi di Milano-Bicocca, Milan, Italy [2] Computational Genomics, IBM T. J. Watson Research Center, Yorktown Heights, USA

§ Corresponding author parida@us.ibm.com

APC is the poster presenter annapaola.carrieri@disco.unimib.it

Abstract

Simulating populations is a fundamental problem in population genetics and is crucial in many applied areas. A generative model simulates the population by evolving a population over time. Here we use the Wright Fisher population model of genetic variation. Backward simulations (primarily based on coalescence [3]) are usually much faster than forward simulations due to the elimination of genetic transmission paths that are not relevant to the samples under study. When genetic exchange events are modeled in addition to the polymorphisms of the duplication model, the resulting network structure is called an ancestral recombination graph (ARG). The input parameters we use are effective populations size N at each generation, recombination rate r, number of extant samples m, mutation rate of single nucleotide polymorphism (SNP) μSNP and mutation rate of short tandem repeats (STR) μSTR. Note that the topology or shape of the ARG is primarily governed by the values of m and r, while the lengths of edges is dictated by the value of N. Finally, the edges of the ARG are annotated by the polymorphisms, whose overall number is determined by the size of the chromosomal segment being simulated.

Such a simulator is provided by [2] and a good exposition of the algorithm is presented in [1], while [6] incorporates other population events such as migration, bottle-necks and so on into the core method. A combinatorial model that uses random graphs is presented in [4]. The paper also discusses an ancestor without ancestry paradox and the reader is directed to a non-technical exposition in [5]. A theorem of the combinatorial model helps in computing closed form approximations of various expected characteristic values of the ARG, such as the depth of the Grand Most Recent Common Ancestor (GMRCA), the variance and the expected number of mutations in the extant sample population. The implications of the paradox leads to a small modification in the simulation algorithm that is implemented in SimRA. We show that SimRA is much faster than the existing methods, sometimes reducing hours of computations down to minutes. Also, the observed ARG characteristic values (depth of GMRCA; sample variation; number of mutation sites) are more tightly spread around the respective expected values, than the other existing methods.

Acknowledgments

This work was carried out while the first author was visiting IBM T J Watson Research Center. The authors are grateful to Filippo Utro and Daniel Platt for very useful discussions and their unwavering support.

References

[1] J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford Press, 2005
[2] R. R. Hudson. *Generating samples under a Wright-Fisher neutral model of genetic variation.* Bioinformatics, 18:337-338, Feb 2002.
[3] J. F. C. Kingman. *On the Genealogy of Large Populations.* Journal of Applied Probability, 19A:2743, 1982.
[4] L. Parida. *Ancestral Recombinations Graph: A reconstructability perspective using random-graphs framework*. Journal of Computational Biology, 17:1345-1350, 2010.
[5] L. Parida. *Graph Model of Coalescence with Recombinations*. Problem Solving Handbook in Computational Biology and Bioinformatics, pages 85-100, 2010.
[6] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M.J. Daly and D. Altshuler. *Calibrating a coalescent simulation of human genome sequence variation*. Genome Res., 15:1576-1583, Nov 2005.

Poster 6

Pegasus: annotation and prediction of oncogenic gene fusion events as a supervised learning task

Francesco Abate[*1], Sakellarios Zairis[*§ 1], Elisa Ficarra [2], Andrea Acquaviva [2], Chris H. Wiggins [3], Veronique Frattini [4], Anna Lasorella [4], Antonio Iavarone [4], Giorgio Inghirami [5], Raul Rabadan [1]

[1] Department of Systems Biology, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032, USA. [2] Department of Control and Computer Engineering, Politecnico di Torino, Torino 10129, Italy. [3] Department of Applied Mathematics, Columbia University, 500 W 120th Street, Mudd 200, MC 4701, New York, NY 10027, USA. [4] Institute for Cancer Genetics, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032, USA. [5] Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, 525 East 68th Street, Starr Pavilion, 715, New York, NY 10065, USA.

[*] Equal author contribution

[§] Corresponding author and poster presenter, siz2102@columbia.edu

Abstract

RNA sequencing offers a genome-wide view of expressed transcripts, uncovering biologically functional genome rearrangements. A possible consequence of genome rearrangement, gene fusions, rose to prominence with the success of imatinib in the treatment of BCR-ABL1 associated cancers. Although several bioinformatics tools are already available for the detection of putative fusion transcripts, candidate event lists are plagued with non-functional read-through events, reverse transcriptase template switching events, incorrect mapping, and other systematic errors. Such lists lack any indication of oncogenic relevance, and they are too large for exhaustive experimental validation. We have designed and implemented a pipeline, Pegasus, for the annotation and prediction of biologically relevant gene fusion candidates. Pegasus provides a common interface for various gene fusion detection tools, reading-frame-aware annotation of preserved/lost functional domains, and data-driven classification of oncogenic potential. More specifically, the domain annotations form a high dimensional feature space in which each fusion transcript is represented as a vector. Leveraging existing databases of oncogenic fusions, we train a gradient boosted ensemble of decision trees to recover a robust predictor of a transcript's oncogenic potential. We demonstrate the effectiveness of Pegasus in predicting new driver gene fusions in 176 RNA-Seq samples of glioblastoma multiforme and 23 cases of anaplastic large cell lymphoma.

The software is freely available at http://sourceforge.net/projects/pegasus-fus

Poster 7

Discovering common intervals in multiple indeterminate strings

Daniel Doerr [1,2§], Jens Stoye [1,2], Katharina Jahn [3]

[1] Genome Informatics, Faculty of Technology, [2] Institute for Bioinformatics, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany [3] Computational Biology Group, Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

[§] Corresponding author and poster presenter ddoerr@cebitec.uni-bielefeld.de

Abstract

We study indeterminate strings, which are sequences over the non-empty elements of $\mathcal{P}(\Sigma)$, the power set of a finite alphabet Σ. Indeterminate strings have applications in computational biology where they can be used to represent chromosomes as ordered sequences of multiple co-localized genomic markers. We distinguish a strict and a weak version of common intervals in indeterminate strings where either the complete set of elements at every interval position, or at least one element, needs to be shared with the other interval. In our proceedings paper of this conference, we introduced the concept of common intervals to indeterminate strings. In this work, we extend this concept from pairwise to simultaneous comparisons of multiple sequences. We present exact algorithms to detect the complete set of common intervals between two or more indeterminate strings for both, weak and strict common intervals.

Poster 8

Bloom Filter Trie - a data structure for pan-genome storage

Guillaume Holley [1§], Roland Wittler [1], Jens Stoye [1]

[1] International Research Training Group 1906/1 and Genome Informatics, Faculty of Technology, Bielefeld University, Germany

[§] Corresponding author gholley@cebitec.uni-bielefeld.de

Abstract

High Throughput Sequencing technologies have become fast and cheap in the past years. As a result, large-scale projects started to sequence tens to several thousands of genomes per species. The concept of pan-genome has therefore emerged. It is composed of two parts: The core genome, which represents the pool of all the genes shared by all the strains of the species, and the dispensable genome, which represents genes not shared by all strains. Graphs exhibit good properties for representing a pan-genome, as they avoid storing redundant information and can represent genomic variants. Unfortunately, existing tools using this representation are slow or memory consuming. We present here a new data structure for storing pan-genomes: The Bloom Filter Trie. This data structure allows to store a pan-genome as an annotated de-Bruijn graph in a memory efficient way. The insertion of new genomes does not degrade the performance of the structure.

Poster 9

Genotyping by sequencing of a collection of Miscanthus spp. Accessions

Manfred Klaas [1§], Paul Cormican [2], Thibauld Michel [1], Susanne Barth [1]

[1] Teagasc Crops Environment and Land Use Programme, Oak Park Crops Research Centre, Carlow, Ireland [2] Teagasc Animal and Bioscience Research Department, Animal & Grassland Research and Innovation Centre, Grange, Dunsany, Co. Meath, Ireland

[§] Corresponding author and poster presenter manfred.klaas@teagasc.ie

Abstract

A genotyping by sequencing (GBS) experiment to genotype our living Miscanthus spp. collection was initiated. This collection includes ~170 accessions of M. sinensis, M. sacchariflorus, M .x giganteus and ~100 hybrids. In addition we analyzed DNAs of several closely related species. Our procedure follows with modifications the protocol from Elshire et al. 2011 (PLoS ONE 6(5): e19379. doi:10.1371). Illumina HiSeq2000 next generation sequencing was applied to generate large numbers of SNPs from sequencing genomic DNA. The SNPs are used as molecular markers to determine the genetic distances between the accessions. To reduce the complexity of the results, the genomic DNA was fragmented by the restriction enzyme PstI. In a first experiment, we used a set of 96 adapters with different barcodes and sequenced in parallel 96 samples. After sequencing on the Illumina HiSeq 2000 platform 139 million reads were obtained. A second set of DNAs was sequenced at higher read depth to gauge required extent of sequence data for our combination of species and restriction enzyme. We present the analysis of population genetics data after coming through the SNP discovery pipeline.

Poster 10

How much are you willing to pay? Selecting costs for reconciliation with duplication and transfers

Han Lai[1,§], Dannie Durand[1]

[1]Biological Science, Carnegie Mellon University, 4400 Fifth Ave, Pittsburgh, PA, 15213
[§] Corresponding author hanlai@andrew.cmu.edu

Abstract

Gene events, including gene duplication, transfer, and loss, are major forces driving the evolution of genetic novelty. Correctly inferring these events is crucial to relating gene evolution to adaptation to ecological change, understanding the process of gene function evolution, and inferring the homology relationships of genes. The history of events in a gene family can be inferred by reconciliation, comparing the gene tree with the corresponding species tree to find the gene events that best explain the incongruence between gene tree and species tree. Previous reconciliation software focused on models of duplication (D) and loss (L) only for eukaryotes or transfer (T) and loss only for prokaryotes. However, recently algorithms have been developed to incorporate all three events because of new evidence suggesting that transfer and duplication are both active process in some species. In DL-only algorithms the solution is unique. In a DTL model, multiple event histories can explain the same tree incongruence because new gene copies can arise either through duplication or through transfer. In parsimony-based reconciliation algorithms, the best history is selected by minimizing weighted sum of duplications, transfers and losses. Therefore, the optimal solution depends on the cost assigned to each event. The important issue is: what cost should the user pick when using these algorithms, so that the event inference from these algorithm is most accurate?

We will present a study of reconciliation with gene trees simulated with various event rates. The simulated trees were reconciled with various event costs and the fraction of correctly inferred event histories was noted. We then analyzed how the relationship between the inferred events and the true events depends on event costs and rates. We also present an analytical framework for summarizing the tradeoff between transfer, duplication, and loss events for a given tree topology. These results provide a foundation for understanding how inferred events change in a parsimony model when event cost varies.

Poster 11

ASTRAL: fast and accurate species tree estimation from gene trees

Siavash Mirarab [1], Rezwana Reaz [1], Md. Shamsuzzoha Bayzid [1], Theo Zimmermann [1,3], M Shel Swenson [2], Tandy Warnow [1,4 §]

[1] Department of Computer Science, The University of Texas at Austin, Austin, TX 78712, USA [2] Departement d'informatique, Ecole Normale Superieure, 45 Rue d'Ulm, F-75230 Paris Cedex 05, France [3] Department of Electrical Engineering, The University of Southern California, Los Angeles, CA 90089, USA [4] Departments of Bioengineering and Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

§ Corresponding author warnow@illinois.edu

MSB poster presenter bayzid@cs.utexas.edu

Abstract

Species trees provide insight into the evolutionary histories of organisms and have application in understanding the mechanisms of evolution, biomolecular function and structure, biodiversity and biogeography, among other things. Estimating species trees is complicated by the fact that gene trees often differ from species trees. One of the most ubiquitous causes for discordance between gene trees and species trees is incomplete lineage sorting (ILS), which is modeled by the multispecies coalescent process. Many methods have been developed to estimate species trees from multiple genes sampled across the genome taking into account ILS, and some of these methods have statistical guarantees under the multi-species coalescent model. However, existing methods are often too computationally intensive for use with large datasets or have been found to have poor accuracy under some realistic conditions. We present ASTRAL, a fast method for estimating species trees from multiple unrooted genes. ASTRAL uses dynamic programing to find the species tree that shares the maximum number of induced quartet trees with input gene trees; the problem is solved exactly for small datasets, and for larger datasets, a constrained version of the problem is solved. ASTRAL is statistically consistent, can run on datasets with hundreds of taxa and thousands of genes and has outstanding accuracy, improving on MP-EST and the population tree from BUCKy (two statistically consistent leading coalescent-based methods). ASTRAL is also more accurate than concatenation using maximum likelihood, except when ILS levels are low or there are too few gene trees. ASTRAL is available in open source form at https://github.com/smirarab/ASTRAL/. ASTRAL is published as part of Bioinformatics special issue for ECCB (2014): http://doi.org/10.1093/bioinformatics/btu462.

Poster 12

PASTA: ultra-large multiple sequence alignment

Siavash Mirarab [1], Nam Nguyen [1,2], and Tandy Warnow [1,2 §]

[1] Department of Computer Science, The University of Texas at Austin, Austin, TX 78712, USA [2] Departments of Bioengineering and Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA

[§] Corresponding author warnow@illinois.edu

SM poster presenter smirarab@utexas.edu

Abstract

PASTA (Practical Alignment using SATé and TrAnsitivity) is a new and highly scalable algorithm for large-scale multiple sequence alignment estimation. PASTA uses a new technique to produce an alignment given a guide tree that enables it to be both highly scalable and very accurate. The design of PASTA is similar to SATé, but uses a fundamentally different approach for merging alignments that uses transitivity to merge a set of pairwise merged sub-alignments. We present a study on biological and simulated data with up to 200,000 sequences, showing that PASTA produces highly accurate alignments, improving on the accuracy of the leading alignment methods on large datasets, and is able to analyze much larger datasets than the current methods. We also show that trees estimated on PASTA alignments are highly accurate – better than SATé trees, but with substantial improvements relative to other methods. Finally, PASTA is very fast, highly parallelizable, and requires relatively little memory. PASTA is available in open source form at https://github.com/smirarab/PASTA/. PASTA has been published in the proceedings of RECOMB 2014 (doi:10.1007/978-3-319-05269-4_15), and an updated version is to appear in JCB.

Poster 13

The effect of tRNA levels on decoding times of mRNA codons

Alexandra Dana[1], Tamir Tuller[1,§]

[1]The Department of Biomedical Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel
[§] Corresponding author tamirtul@post.tau.ac.il

Abstract

The possible effect of tRNA concentrations on codons decoding time is a fundamental biomedical research question; however, due to a large number of variables affecting this process and the non-direct relation between them, a conclusive answer to this question has eluded so far researchers in the field.

First, gene expression is affected by a large number of factors such as mRNA folding, context of the start codon, charge of the amino acids, molecules concentrations of ribosomes, mRNA, tRNA and Aminoacyl tRNA synthetases; thus it is impossible to completely control for non-causal relations between these two variables. Second, heterologous gene expression, a common method for studying such relations, may not reflect the decoding time of endogenous transcripts since they tend to violate the natural intracellular regimes. Third, most large scale gene expression measurements do not directly measure translation elongation rates (*e.g.* protein levels).

The current cutting edge methodology for studying mRNA translation is ribosome profiling. This deep-sequencing method produces a detailed account of ribosome occupancy on specific mRNAs. Recently, several studies analyzing this data found insignificant correlations between tRNA levels and codons decoding times, inconsistent with previous studies based on other methodologies and data sources.

In this study we perform a novel analysis of the ribosome profiling data of four organisms that enables ranking the decoding times of different codons. We show for the first time that when filtering out rare events such as long translational pauses and the influence of ribosome traffic jams, the correlation between codon decoding times and the tRNA concentrations is significant in all analysed organisms (-0.38 to -0.66, all p values < 0.006). In addition, we show that when considering tRNA concentrations, codons decoding times are not correlated with aminoacyl-tRNA levels.

This finding should help to understand the evolution of synonymous aspects of coding sequences via the adaptation of their codons to the tRNA pool. In addition, this relationship is not only fundamental for human health, but also affects biotechnology and disciplines such as molecular evolution and functional genomics.

Poster 14

CompPSA: A Component-Based Pairwise RNA Secondary Structure Alignment Algorithm

Ghada Badr[1*§], Arwa AlTurki[1*]

[1]Computer Sciences Department, King Saud University, Riyadh, Saudi Arabia
[*]These authors contributed equally to this work
[§]Corresponding author and poster presenter: badrghada@hotmail.com

Abstract

The function of an RNA molecule depends on its structure. The objective of the alignment is finding the homology between two or more RNA secondary structures. Knowing the common functionalities between two RNA structures allows a better understanding and a discovery of other relationships between them. Besides, identifying non-coding RNAs -that is not translated into a protein- is a popular application in which RNA structural alignment is the first step.

Some methods for RNA structure-to-structure alignment have been developed. Most of these methods are partial structure-to-structure, sequence-to-structure or structure-to-sequence alignment. In this work, we introduce an $O(N^2)$ Component-based Pair-wise RNA Structure Alignment (CompPSA) algorithm,. N is the maximum number of components in the two structures, where Structures are given as a component-based representation. The proposed algorithm compares the two RNA secondary structures based on their weighted component features rather than their base-pair details. In addition, we propose similarity measures between two RNA secondary structures so that they can efficiently reflect the similarity between their components.

Extensive experiments are conducted illustrating the efficiency of the CompPSA algorithm when compared to other approaches and on different datasets. The CompPSA algorithm shows an accurate similarity measure between components. The algorithm gives the flexibility for the user to align the two RNA structures based on their weighted features (position, full length, and/or stem length). Moreover, the algorithm proves scalability and efficiency in time and memory performance.

Acknowledgments

This research is supported by the National Plan for Sciences and Technology, King Saud University, Riyadh, Saudi Arabia (Project No. 12-BIO2605-02).

Poster 15

Single-cell sequencing: How many is many enough?

Robert Aboukhalil [1§], Joan Alexander [1], Jude Kendall [1], Michael Wigler [1], Gurinder S. Atwal [1]

[1] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
§ Corresponding author raboukha@cshl.edu

Abstract

Recent developments have paved the way for DNA sequencing at single-cell resolution. These approaches enable the precise study of tumor heterogeneity, circulating tumor cells, neuronal mosaicism, and other complex biological phenomena. Despite these advances, the question of how many cells to sequence in order to capture a significant portion of a sample's heterogeneity remains unanswered. To address this question within the context of copy-number alterations, we performed simulations on several unpublished single-cell datasets from prostate tumor biopsies (over 350 cells per patient). For each patient sample, we constructed a phylogenetic tree from copy-number variations between the cells. Using hundreds of down-sampling simulations, we identified the minimum number of cells needed to capture the clonal structure of the sample, using metrics such as the major number of clusters and the conservation of cluster content. We show that there exists a critical number of cells below which analysis is swamped by noise and fails to capture the heterogeneity of the sample. For the samples we analyzed, this phase transition happens at ~100 cells, although we expect this number to be sample-specific. Next, we explored the tradeoff between sequencing more cells and sequencing at higher depth. We obtained several single-cell datasets—both unpublished and published—from breast tumors, prostate cancer biopsies and neurons. By simulating the effects of using different combinations of number of cells and number of reads per cell, we can estimate how accurately we would have reconstructed the tree in these conditions. For most samples, we find that to reveal clonal substructure, sequencing more cells is more effective than sequencing at higher depth. However, to assess the degree of genomic instability, greater depth of coverage is valuable. Both can be accomplished: first an inexpensive survey of all cells at low coverage depth, then selecting a subset of cells for greater depth of sequencing.

Poster 16

Evolution of genes neighborhood within reconciled phylogenies: an ensembl approach

Cedric Chauve[1 §], Yann Ponty[1, 2], João Paulo Pereira Zanetti[1, 3, 4]

[1] Department of Mathematics, Simon Fraser University, Burnaby, Canada [2] Pacific Institute for Mathematical Sciences, CNRS UMI3069, Vancouver, Canada [3] Institute of Computing, UNICAMP, Campinas, Brazil [4] São Paulo Research Foundation, FAPESP, São Paulo, Brazil

[§] Corresponding author cedric.chauve@sfu.ca

Abstract

Reconstruction of the evolutionary history of genomic characters along a given species tree is a long-standing problem in computational biology, with efficient algorithms to compute parsimonious scenarios for many types of characters, as in the cases of genes and genomes sequences, gene content, and gene family evolution. Recently, Bérard et al. extended the corpus of such results to synthetic characters, as they introduced the notion of adjacency forest, that models the evolution of gene adjacencies within a phylogeny, and described an efficient dynamic programming (DP) algorithm, called DeCo (Berard et al, ECCB 2012), to compute parsimonious adjacency evolutionary histories.

Applying the classical parsimony-based approach of DeCo to a dataset of over 6,000 pairs of mammalian gene trees yielded a significant number of ancestral genes involved in more than two adjacencies, which correspond to synthetic inconsistencies.

Recently, more general approaches for parsimony problems have been analyzed, either exploring a wider range of parameters, or considering several alternate histories for a given instance.

Our work seeks to address the synthetic inconsistencies of DeCo by extending their DP scheme toward an exploration of the whole solution space of adjacency histories, under the Boltzmann probability distribution, that assigns a probability to each solution defined in terms of its parsimony score. This principle has been applied in several contexts and is sometimes known as the "Boltzmann ensemble approach".

While this Boltzmann ensemble approach has been used for a long time in RNA structure analysis, to the best of our knowledge it is not the case in comparative genomics, where exact probabilistic models have been favored as increasing computational capacities allow them to handle realistic datasets. However, such a probabilistic model does not exist so far for gene adjacencies, which motivates our work.

We first show that by sampling adjacencies histories under a Boltzmann distribution that favors co-optimal histories and conserving only frequent ancestral adjacencies, we can reduce significantly the number of synthetic inconsistencies. We also implement a inside/outside variant for DeCo, that computes the actual probabilities of each individual adjacency considered under the Boltzmann distribution.

Poster 17

Genome and sequence characteristics indicate frequent introgressive hybridization events in monocots and dicots

Sapna Sharma[1], Thomas Nussbaumer[1], Karl Kugler[1], Klaus F. X. Mayer [1,§]

[1]Plant Genome and Systems Biology, Helmholtz Center Munich,

[§] Corresponding author k.mayer@helmholtz-muenchen.de

SS is the poster presenter sapna.sharma@helmholtz-muenchen.de

Abstract

With the availability of a series of finished plant genomes detailed insights into the degree and extent of syntenic conservation among different plant genomes became possible. We investigated similarities and dissimilarities among corresponding syntenic regions in a broad variety of different monocot and dicot genomes. Using CDS we identify pairs of orthologous genes among barley (*Hordeum vulgare*) with respect to Brachypodium (*Brachypodium distachyon*), Sorghum (*Sorghum bicolor*) and rice (*Oryza sativa*). For dicots, *Arabidopsis thaliana* was used as reference and compared against *Arabidopsis lyrata* and *Brassica rapa*. Similarity comparisons and substitution rates show that significant differences among syntenic segments are apparent. We interpret these findings as indicative of frequent and repeated introgressive hybridization during the evolution of the individual genomes. Beside whole genome duplications this mechanism might be important for the genome evolution of plants.

Poster 18

Scaffolding of Ancient Contigs and Ancestral Reconstruction in a Phylogenetic Framework

Nina Luhmann [1§], Cedric Chauve [2], Jens Stoye [1], Roland Wittler [1]

[1] International Research Training Group "Computational Methods for the Analysis of the Diversity and Dynamics of Genomes" and Genome Informatics, Faculty of Technology and Center for Biotechnology, Bielefeld University, Germany [2] Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada

§ Corresponding author nluhmann@techfak.uni-bielefeld.de

Abstract

The knowledge about the structure of ancient genomes can shed light on the evolutionary processes underlying the development of extant genomes. Recent progress in sequencing ancient DNA found in conserved remains allows the integration of this sequencing data in genome evolution analysis. However, the assembly of ancient genomes is fragmented because of DNA degradation over time, resulting in a challenging scaffolding step. We address the issue of genome fragmentation in the assembly within a phylogenetic framework while improving the reconstruction of all ancient genomes in the phylogeny. The tree is augmented with the fragmented assembly of the ancient genome represented as an assembly graph indicating a conflicting ordering of contigs in the assembly.

Our approach is to compare the ancient data with extant related genomes and to reconstruct genomes that minimize the Single-Cut-or-Join rearrangement distance [1] along the tree. In contrast to most rearrangement distances, Feijão and Meidanis showed that this minimization can be computed in polynomial time. However, the result is only consistent if the input data is conflict-free.

We generalize the reconstruction approach minimizing the SCJ distance towards multifurcating trees with the Hartigan algorithm [2] and include edge lengths to avoid a sparse reconstruction in practice. Although the ancient DNA data is not conflict-free, we present an approach to include the additional data in the reconstruction while still ensuring consistent results in polynomial time.

References

[1] P. Feijão and J. Meidanis. SCJ: a breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(5):1318–1329, 2011.
[2] J. A. Hartigan. Minimum mutation fits to a given tree. Biometrics, 53–65, 1973.

Poster 19

Genome Rearrangement for RNA Secondary Structure Using a Component-Based Representation: An Initial Framework

Ghada Badr[1]*[§], Haifaa Alaqel[1]*

[1] King Saud University, College of Computer and Information Sciences, Computer Science Department, Riyadh, Kingdom of Saudi Arabia

*These authors contributed equally to this work

[§]Corresponding author and poster presenter badrghada@hotmail.com

Abstract

Genome rearrangements are essential processes for evolution and are responsible for the existing varieties of genome architectures. Many studies have been made to obtain an algorithm that can find the minimum number of inversions that are necessary to transform one genome into the other for a sequence representation of a genome in a polynomial time. Up to our knowledge, no studies have been made to rearrange the genome when represented as a secondary structure. Unlike sequences, the secondary structure preserves the functionality of the genome. Sequences can be different, but still sharing the same structure and hence the same functionality.

Given two different RNA structures A and B, represented using a recently proposed component-based representation; we propose a new framework that can detect the minimum number of events that make A more or exactly similar to B. It can also report one scenario of these events. We propose genome rearrangement operations for RNA secondary structure (insertion, deletion, reversal, transposition, exchange) and device a novel algorithm to calculate the minimum number of rearrangement operations that are required to transform one structure into the other.

The algorithm is able to calculate the distance between two RNA secondary structures and report at least one scenario that is based on the minimum rearrangement operations that are required to make the given structure more similar (or exact).

The operations and algorithms that are proposed here allow describing for the first time evolutionary scenarios that are based on secondary structures rather than on sequences. This represents a good start for a framework for applying rearrangement operations on secondary structures rather than just sequences, which can help in more understanding the common functionalities between different species.

Acknowledgments

This research is supported by the National Plan for Sciences and Technology, King Saud University, Riyadh, Saudi Arabia (Project No. 12-BIO2605-02).

Poster 20

Identifying risk-associated regulatory SNPs in ChIP-seq enhancers

Di Huang[1] and Ivan Ovcharenko[1§]

[1]Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20892, USA

[§]Corresponding author ovcharen@nih.gov

DH is the poster presenter di.huang@nih.gov

Abstract

SNPs map predominantly to the noncoding regions of the human genome. Although thousands of non-coding SNPs have been linked to human diseases in the past, prioritizing this huge pool of non-coding SNPs is largely impossible due to the inability to accurately quantify the impact of non-coding variation. To overcome this challenge, we developed a computational model that uses ChIP-seq intensity variation in response to non-coding allelic change as a proxy to the quantification of the biological role of non-coding SNPs. We applied this model to HepG2 enhancers and detected 4,796 enhancer SNPs capable of disrupting enhancer activity upon allelic change. These SNPs are significantly overrepresented in the binding sites of HNF4 and FOXA families of liver transcription factors and liver eQTLs. In addition, these SNPs are strongly associated with liver GWAS traits, including type I diabetes, and are linked to the abnormal levels of HDL and LDL cholesterol. Our model is directly applicable to any enhancer set to prioritize non-coding SNPs.

Poster 21

Characterizing Horizontal Gene Transfer in Microbial Evolution using Topological Data Analysis

Kevin Emmett [1,2§], Raul Rabadan [2,3]

[1] Department of Physics, Columbia University [2] Department of Systems Biology, Columbia University

[3] Department of Biomedical Informatics, Columbia University

[§] Corresponding author kje2109@columbia.edu

Abstract

Mounting evidence for the role of horizontal gene transfer in microorganism evolution has led to the need for alternative representations of evolutionary relationships that do not presuppose treelike topology. We apply methods from topological data analysis (TDA) to this problem. These methods have recently been shown to efficiently quantify horizontal evolution in genomic data. Specifically, by representing genomic data in a high-dimensional space equipped with a suitable metric, simplicial complexes can be constructed and parameterized with an evolutionary scale parameter. We find that topological invariants computed by persistent homology on these complexes reveals meaningful biological information at multiple scales. In particular, these invariants can be used to estimate the scale and frequency of horizontal events in genomic data, as well as population structure. We use this approach to estimate the extent of horizontal gene transfer in gene family presence/absence data from prokaryotes and bacteriophages.

Compared to network reconstruction methods, persistent homology does not depend on a specific scale threshold, but rather contains quantitative information about all scales.

Poster 22

Role of p33ING1b in Head and Neck Cancer

Mehmet Gunduz[1§], Esra Gunduz[1], Omer Faruk Hatipoglu[1], Gokhan Nas[1], Elif Nihan Cetin[1], Bunyamin Isik[2] , M.Ramazan Yigitoglu[3]

[1]Department of Medical Genetics, [3]Department of Medical Biochemistry, Faculty of Medicine, Turgut Ozal University, Ankara, Turkey [2] Department of Family Medicine, Faculty of Medicine, Hacettepe University, Ankara, Turkey

[§] Corresponding author mehmet.gunduz@gmail.com

BI is the poster presenter

Abstract

3-5% of all cancer cases consist of head and neck cancer. Survival rates changes according to its grade and it is a severe threat to human life. Currently surgery is the main treatment modality and it leads to major deformities in head and neck area. Moreover important functions such as swallowing and voice is disturbed and this causes to decrease in quality of life. Thus researches involving genetic basis of head and neck cancer and alternative treatment methods based on these studies are warranted.

We characterized ING1 genomic structure and showed its tumor suppressive role in head and neck cancer for the first time in the literature. It has three exons and four introns and two major splicing variants of p24ING1c and p33ING1b. In the current study we aimed to clarify the role p33ING1b in carcinogenesis and metastasis of head and neck cancer. We have constructed p33ING1b expression plasmids to head and neck cancer cell lines and examined mRNA and protein levels. Cell proliferation, migration and invasion assays using p33ING1b plasmid in head and neck cancer cell lines were performed. Moreover we've performed the cell cycle analysis to detect at which step p33ING1b stops cell cycle. And we've confirmed our results by overexpression of p33ING1b through siRNA experiments.

The results showed that p33ING1b inhibited cell proliferation, migration and invasion in head and neck cancer cell lines. Also when p33ING1b is supressed with siRNA, cell proliferation is accelerated. Cell cycle analysis displayed that p33ING1b stops cell cycle at G1. We've shown p33ING1b's important role in head and neck cancer through our experimental results and to the best of our knowledge this is the first study which shows p33ING1b's role in head and neck cancer.

Poster 23

The Family-Free Double Cut and Join and its application to ortholog detection

Pedro Feijao [1§], Fábio V Martinez [1,2], Marília Braga [3], Jens Stoye [1]

[1] Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany. [2] Universidade Federal do Mato Grosso do Sul, Cidade Universitaria, MS, 79090-900, Brazil. [3] Bioinformatics research group, INMETRO, Rio de Janeiro, Brazil

[§] Corresponding author and poster presenter pfeijao@cebitec.uni-bielefeld.de

Abstract

In the course of evolution, genomes are subject to large-scale mutations and rearrangements, such as inversions, translocations and duplications of large blocks of DNA. A classical problem is to compute the rearrangement distance, that is, the minimum number of rearrangement events required to transform a given genome into another. A frequently used model to solve this problem is the Double-Cut-and-Join (DCJ) operation, a simple and comprehensive operation that mimics several rearrangements.

A usual pre-processing step in several comparative genomics studies is the classification of genes in all genomes into gene families, where genes in the same family are said to be homologous. We call this approach ``family-based". This classification is commonly done automatically and is subject to errors and loss of information that can potentially compromise the results of the subsequent analysis. Due to this fact, an alternative ``family-free'' approach was recently proposed, where no prior family assignment is made and only the pairwise similarity between genes is used as input.

The authors recently proposed the Family-Free DCJ distance problem (FFDCJ), that is, finding the DCJ distance between two genomes, given the pairwise similarity between all genes as input. This problem was shown to be NP-hard and an ILP was proposed to solve it. In this work, we show some improvements and variations and the original ILP and solve the ILP on simulated datasets, demonstrating how the FFDCJ can be used for phylogeny reconstruction and orthology detection.

Poster 24

LearnedPhyloblocks: Novel Genomic Islands through Phylogenetic Profiling

Corey M. Hudson[1§], Kelly P. Williams[1]

[1] Sandia National Laboratories, Department of Systems Biology, Livermore,CA, USA

[§] Corresponding author and poster presenter cmhudso@sandia.gov

Abstract

Mobile genomic elements, capable of moving genes between prokaryotes, have a major role in determining the pathogenicity of bacterial strains. These include genomic islands, which move large segments of DNA, integrons and plasmids. Genomic islands, in particular, have a major role in pathogenicity, having been implicated in the shift from nonpathogenic to pathogenic strains of Vibrio, Yersinia, enterohaemorrhagic Escherichia coli, Burkholderia cenocepacia, and MRSA, to name a few. We have developed the Islander database and software suite to identify and characterize high-quality genomic islands from raw nucleotide sequence. Our algorithm uses established markers of genomic islands, including tRNAs, integrases and tRNA fragments to identify mechanically determined genomic islands. This method serves as the training set for probabilistically determining other recently transferred genomic regions. This technique involves the collection of phylogenetic profiles for every position in the genome, and then implements a supervised learning strategy to determine the proportion of given phylogenetic profiles that are found in and out of known islands. It then uses the Islander database to identify patterns of island enrichment in blocks of contiguously aligned sequence. Using the highly antibiotic resistant Klebsiella pneumonia BAA-2146 as a test case, we used previously identified genomic islands in this genome as a training set for probabilistically determined recently transferred genomic regions, using phylogenetic profiles across Klebsiella. We found a new genomic island (carrying a resistance integron), the capsular polysaccharide synthesis region – a primary pathogenicity determinant and four insertion sequences, two of which confer resistance to antibiotics. We have implemented this algorithm as a software tool LearnedPhyloblocks, which can identify novel transferred elements, using the Islander software package and genome alignments across any given genera.

Poster 25

Tracking the Evolution of a Signal Transduction Pathway Architecture with Comparative Genomics

Philip Davidson [1§], N Luisa Hiller [1], Michael T Laub [2], Dannie Durand [1]

[1] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [2] Department of Biology, Massachusetts Institute of Technology, Boston, MA 02139, USA; Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
[§] Corresponding author pdavidso@andrew.cmu.edu

Abstract

Endosporulation is a key characteristic of the Firmicutes bacteria, however, the Sporulation Initiation pathway in the two most extensively studied genera have two different pathway architectures. The Clostridial pathway is a simple two component system, while the Bacilli have four proteins that interact in a phosphorelay architecture. Despite the differences in architecture these pathways share many characteristics: the terminal proteins in both pathways are extremely similar in sequence, the proteins involved in the rest of each pathway consist of conserved domains, and both pathways activate analogous cellular differentiation programs. These observations, taken together, support the hypothesis that both pathways are derived from the same pathway in the Firmicute ancestor. This is an excellent case study for network architecture evolution and rewiring. A Firmicute-wide survey for Sporulation Initiation pathway components is a prerequisite to reconstructing the evolutionary trajectory of these pathways. To perform this survey we relied heavily on genome neighborhoods to identify candidate components and used similarity in the set of amino acids that control interaction specificity to validate candidate components. The phylogenetic distribution of candidate components supports the hypothesis that the ancestral state consisted of four components in a phosphorelay architecture. Our results challenge the widely accepted pathway evolution hypothesis that the complex phosphorelay architecture resulted from expansion and elaboration of the simple two component system.

Poster 26

Linking Genome Rearrangements and Chromatin Conformation

Krister M. Swenson [1§], Mathieu Blanchette [2]

[1] Institut de Biologie Computationnelle, Méthodes et algorithmes pour la bioinformatique, LIRMM, Montpellier and CNRS, France [2] McGill University

[§] Corresponding author swenson@lirmm.fr

Abstract

Large-scale rearrangements drastically change linear gene orders. These moves are significant since proximity on the linear gene order is linked to gene co-expression and co-regulation in many species across the tree of life, including in human. Rearrangements inhibit subsequent crossover so are thought to increase genetic variability, and are a mechanism for enforcing reproductive isolation. Thus the primary mechanisms and constraints governing the advent and fixation of rearrangements in a population are of high interest.

A clean and simple picture of genotype evolution is consistent with the hypothesis that evolutionarily conserved rearrangements happened between pairs of breakpoints that were generally close in 3D space; normal cell function would not have been greatly disturbed, and known mechanisms explain the seemingly large-scale change. Cancer-causing somatic rearrangements seem to support this hypothesis. The advent of Hi-C methods for chromosome capture has recently opened the door to similar study on an evolutionary scale. Yaffe *et al.* showed that rearrangement breakpoint pairs between human and mouse are concentrated around replicating domains, and occur at locations with similar time-of-replication. They also showed that some subset of 55 interchromosomal breakpoint pairs existing in human (with respect to mouse) correlate with interaction frequency. Veron *et al.* took this result further by studying many more of the breakpoint pairs between human and mouse. They found a significant correlation between 3D proximity, and intrachromosomal breakpoint pairs. They found no such correlation for interchromosomal pairs.

In this poster, we present new results showing a strong correlation between evolutionary breakpoints of rearrangements and spacial proximity in the nucleus of human. The pattern exists for intra and interchromosomal pairs, and is consistent across multiple cell lines from multiple labs. We show this by sampling DCJ rearrangement scenarios between human and mouse, and comparing the chromosome proximity on the scenarios to a null hypothesis on random breakpoint locations. We then present new optimization problems related to finding ``local'' rearrangement scenarios, and give preliminary algorithmic results.

Poster 27

Topological Data Analysis to detect population admixture in recombining chromosomes

Filippo Utro [1], Deniz Yorukoglu [2], David Kuhn [3], Saugata Basu [4], Laxmi Parida [1§]

[1] IBM T. J. Watson Research, New York, USA. [2] MIT, Massachusetts, USA. [3] USDA-ARS, Florida, USA. [4] Purdue University, Indiana, USA.
§ Corresponding author parida@us.ibm.com
FU is the poster presenter futro@us.ibm.com

Abstract

Diploid populations undergo recombinations at each generation and as a result, a collection of extant chromosomes show a mosaic pattern of chromosomal segments. The evolution history of the extant individuals, of different populations is captured in the network called an ancestral recombinations graph (ARG). Note that ARG is rampant with closed paths due to recombinations. Then how does the ARG capture mixing of populations? We model this as a *scaffold* (a network structure) with the extant populations at the leaf nodes. Each edge of this scaffold represents an evolving population where each junction of the scaffold represents either ‘‘diverging’’ or ‘‘converging’’ populations. If the scaffold has no junction with multiple parents, then the resulting populations are not admixed. Based on the model above results from topology, we prove the following. We have a set M (set of individuals) with a distance measure d, and a surjective map $Pop: M \rightarrow M'$ that clusters individuals into populations. On M' there is an induced distance d' defined as the minimum distance between individuals of the two populations. For each t, we have a Rips complex $Rips(M,t)$ whose homology we can calculate. We are interested in knowing if there is any persistent homology in the Rips complex of M' (of the populations). This indicates presence of admixtures. We prove, under certain natural restrictions on the distance measure d,d':

> *Theorem: For any persistent homology cycle [c] in the Rips complex of M′, there exists a unique persistent homology cycle in the Rips complex of M that maps to [c] (by a homomorphism induced by the map Pop).*

Based on this theorem, we carry out simulation experiments to use topological signatures to detect admixture. The Vietoris-Rips complex was constructed on the graph embedding of the distance matrix (a complete graph with each vertex corresponding to an individual haplotype and edge weights corresponding to the Hamming distance between the pair of haplotypes). We computed homology groups on the Vietoris-Rips complex for zero and one dimension. Then we applied this to avocado germplasm data to detect admixture. Our preliminary results of using topological structures as signatures for admixture look promising.

Acknowledgments

The work was done while DY was a summer intern at IBM T. J. Watson Research Center. We thank Anna Paola Carrieri for implementing the software that was used in experiments for the simulations of the populations.

Poster 28

Achieving Cross-Platform Compatibility of Gene Expression Data

Yee Him Cheung[1], Nevenka Dimitrova[1], Wim Verhaegh[2§]

[1] Clinical Informatics Services and Solutions Department, Philips Research North America, Briarcliff Manor, NY, USA [2] Precision and Decentralized Diagnostics Department, Philips Research, Eindhoven, Netherlands

[§]Corresponding author wim.verhaegh@philips.com

YHC is the poster presenter patrick.cheung@philips.com

Abstract

The dynamic ranges of gene expressions can vary considerably depending on the choice of profiling platform. As a result, prognostic gene signatures are usually platform-specific, and expression data generated by heterogeneous platforms in general cannot be directly combined for computational analysis, thus limiting the scope of use of legacy data and hindering the adoption of new profiling technologies. Specifically, tremendous resources have been spent on microarray studies, and it is desirable to transfer the knowledge and insights onto the new platform such as next generation sequencing. Towards the goal of cross-platform compatibility of gene expression data, we developed regression-based models that transform expression data from one platform to another. To preserve the predictive power of a gene signature on alternative platforms, we implemented an algorithm that selects a subset of gene targets based on criteria such as their individual statistical powers in distinguishing the subtypes, their detectability as indicated by the average expression levels and their concordance with the primary platform in terms of their relative abundance across subtypes. We in particular investigated the compatibility between Illumina HiSeq 2000 RNA-Seq (log2 RSEM) and Affymetrix HT-HG-U133A microarray (log2 RMA) expressions using 545 TCGA samples that were studied on both platforms. On average, the correlation between the two platforms is $r = 0.71$ per sample, and for moderate-to-high microarray expressions, the predicted RNA-Seq expressions have a root mean square error $e_{rms} = 1.4$, which is very close to that of 1.39 obtained by direct regression. By suitably selecting a subset of targets to represent a gene signature derived from microarray expressions on the RNA-Seq platform, our proposed algorithm showed promising performance for pathway activity prediction based on a Bayesian approach, and successfully adapt the signature for use on the new platform.

Poster 29

K-mer Analysis of Ebola sequences differentiates outbreaks

Filippo Utro [1], Daniel E. Platt [1], Laxmi Parida [1§]

[1] IBM T. J. Watson Research, New York, USA.
§ Corresponding author parida@us.ibm.com
FU is the poster presenter futro@us.ibm.com

Abstract

In the last few months, the current Ebola outbreak's size and expansion has become a major news story, and its understanding is currently a subject of importance, even though the first documented case on primates was recorded already 45 years ago. The Ebola virus has heretofore been associated with small but alarming outbreaks of hemorrhagic fever in human and nonhuman primates.

Similarity of sequences is a key mathematical notion for classification and phylogenetic studies in biology. We are, to the best of our knowledge, the first to classify the Ebola virus genomes using alignment-free techniques. In this work, we are studying the intrinsic structure of several Ebola virus genome sequences, in terms of *k*-mers. In particular, in this stage we focus on the case $k=3$, obtaining significant results.

We note that there is a direct relationship between Manhattan distances between *k*-mer frequency counts and genetic distances. Moreover, in order to understand how *k*-mer frequencies predict phylogeny in an alignment-free setting, we seek to quantify how much information regarding alignment is actually carried by each *k*-mer. We therefore seek to quantify how many *k*-mers in a sequence might be expected by chance given the relative frequency distribution of *k*-mers and their alignment frequencies. To that end, we will extend the formalism of Karlin-Altshul scoring to determine the distribution of alignment lengths expected by chance given an alignment score. This provides a context for establishing how many *k*-mer differences would be required to recognize both the genetic distance under alignment, as well as *k*-mer frequency distances as a basis for phylogenetic construction.

Our analysis is able to differentiate Ebola outbreaks and provide interesting features in them that merit future analysis. Next, we plan to investigate additional capabilities of our methodology, and extend it to higher values of *k* as well as different organisms.