Short Abstracts

SLICER: Inferring Branched, Nonlinear Cellular Trajectories from Single Cell RNA-seq Data

Joshua D. Welch[1], Ziqing Liu[2], Li Wang[2], Junjie Lu[3], Paul Lerou[3], Jeremy Purvis[4], Li Qian[2], Alexander Hartemink[5], and Jan F. Prins[1]

[1] Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, USA
{jwelch,prins}@cs.unc.edu
[2] Department of Pathology, The University of North Carolina at Chapel Hill, Chapel Hill, USA
[3] Department of Pediatric Newborn Medicine, Harvard Medical School, Boston, USA
[4] Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, USA
[5] Department of Computer Science, Duke University, Durham, USA

1 Abstract

Understanding the dynamic regulation of gene expression in cells requires the study of important temporal processes, such as differentiation, the cell division cycle, or tumorigenesis. However, in such cases, the precise sequence of changes is generally not known, few if any marker genes are available, and individual cells may proceed through the process at different rates. These factors make it very difficult to judge a given cell's position within the process. Additionally, bulk RNA-seq data may blur aspects of the process because cells at sampled at a given wallclock time may be at differing points along the process. The advent of single cell RNA-seq enables study of sequential gene expression changes by providing a set of time slices or "snapshots" from individual moments in the process. To combine these snapshots into a coherent picture, we need to infer an "internal clock" that tells, for each cell, where it is in the process.

Several techniques, most notably Monocle and Wanderlust, have recently been developed to address this problem. Monocle and Wanderlust have both been successfully applied to reveal biological insights about cells moving through a biological process. However, a number of aspects of the trajectory construction problem remain unexplored. For example, both Monocle and Wanderlust assume that the set of expression values they receive as input have been curated in some way using biological prior knowledge. Wanderlust was designed to work on data from protein marker expression, a situation in which the number of markers is relatively small (dozens, not hundreds of markers) and the markers are hand-picked based on prior knowledge of their involvement in the process. In the initial application of Monocle, genes were selected based on differential expression analysis of bulk RNA-seq data collected at initial and final timepoints. In addition, Monocle uses ICA, which assumes that the trajectory lies along a linear projection of the data. In general, this linearity assumption may

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 239–240, 2016.
DOI: 10.1007/978-3-319-31957-5

not hold in biological systems. In contrast, Wanderlust can capture nonlinear trajectories, but works in the original high-dimensional space, which may make it more susceptible to noise, particularly when given thousands of genes, many of which are unrelated to the process being studied. Another challenging aspect of trajectory construction is the detection of branches. For example, a developmental process may give rise to multiple cell fates, leading to a bifurcation in the manifold describing the process. Wanderlust assumes that the process is non-branching when constructing a trajectory. Monocle provides the capability of dividing a trajectory into a branches, but requires the user to specify the number of branches.

In this paper, we present SLICER (Selective Locally linear Inference of Cellular Expression Relationships), a new approach that uses locally linear embedding (LLE) to reconstruct cellular trajectories. SLICER provides four significant advantages over existing methods for inferring cellular trajectories: (1) the ability to automatically select genes to use in building a cellular trajectory with no need for biological prior knowledge; (2) use of locally linear embedding, a nonlinear dimensionality reduction algorithm, for capturing highly nonlinear relationships between gene expression levels and progression through a process; (3) automatic detection of the number and location of branches in a cellular trajectory using a novel metric called geodesic entropy; and (4) the capability to detect types of features in a trajectory such as "bubbles" that no existing method can detect. Comparisons using synthetic data show that SLICER outperforms existing methods, particularly when given input that includes genes unrelated to the trajectory. We demonstrate the effectiveness of SLICER on newly generated single cell RNA-seq data from human embryonic stem cells and murine induced cardiomyocytes.

Multi-track Modeling for Genome-Scale Reconstruction of 3D Chromatin Structure from Hi-C Data

Chenchen Zou[1,8], Yuping Zhang[2,3,4,5,8], and Zhengqing Ouyang[1,3,6,7,8]

[1] The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
zhengqing.ouyang@jax.org
[2] Department of Statistics, University of Connecticut, Storrs, CT, USA
[3] Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA
[4] Institute for Collaboration on Health, Intervention, and Policy, University of Connecticut, Storrs, CT, USA
[5] Center for Quantitative Medicine, University of Connecticut Health Center, Farmington, CT, USA
[6] The Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, Storrs, CT, USA
[7] Department of Biomedical Engineering, University of Connecticut, Storrs, CT, USA
[8] Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT, USA

Abstract. Genome-wide chromosome conformation capture (Hi-C) has been widely used to study chromatin interactions and the 3D structures of the genome. However, few computational approaches are existing to quantitatively analyze Hi-C data, thus hindering the investigation of the association between 3D chromatin structure and genome function. Here, we present HSA, a novel approach to reconstruct 3D chromatin structures at the genome-scale by modeling multi-track Hi-C data. HSA models chromatin as a Markov chain under a generalized linear model framework, and uses simulated annealing to globally search for the latent structure underlying the cleavage footprints of different restriction enzymes. HSA is robust, accurate, and outperforms or rivals existing computational tools when evaluated on simulated and real datasets in diverse cell types.

Keywords: 3D chromatin structure · Multi-track modeling · Genome-wide · Hi-C

M. Singh (Ed.): RECOMB 2016, LNBI 9649, p. 241, 2016.
DOI: 10.1007/978-3-319-31957-5

Revealing the Genetic Basis of Immune Traits in the Absence of Experimental Immunophenotyping

Yael Steuerman and Irit Gat-Viks

Department of Cell Research and Immunology, The George S. Wise Faculty of Life Sciences, Tel Aviv University, 6997801 Tel Aviv, Israel
iritgv@post.tau.ac.il

Introduction and Motivation. The immune system consists of hundreds of immune cell types working coordinately to maintain tissue homeostasis. Thus, discovering the genetic control underlying inter-individual variation in the abundance of immune cell subpopulations requires simultaneous quantification of numerous immune cell types. Current experimental technologies, such as fluorescence-activated cell sorting (FACS) [1], can follow the dynamics of only a limited number of cell types, hence hindering a comprehensive analysis of the full genetic complexity of immune cell quantities. One possible way to attain a global immunophenotyping is to mathematically infer, by means of a deconvolution technique [2–5], the abundance of a variety of immune cell subpopulations based on gene-expression profiles from a complex tissue, without the need of direct cell sorting measurements. Based on these predicted immunophenotypes, a genome-wide association study can be applied to uncover the genetic basis for these immune traits.

Methods. We developed a novel computational methodology to identify significant associations between immune traits and polymorphic DNA loci. Our method combines (i) prior knowledge on the transcriptional profile of various immune cell-types; (ii) gene-expression data in a given cohort of individuals; and (iii) genotyping data of the same individuals. Our method utilizes a deconvolution method which computationally infers the global dynamics of immune cell subsets for each individual. Specifically, we exploit associations between cell types, genes and genotypes to select an informative group of marker genes, rather than the full transcriptional profile, to attain a more accurate deconvolution-based model.

Results. We applied our method to both synthetic and real biological data to evaluate its ability to uncover the genetic basis of immune traits. Our analysis of synthetic data confirms that our method can handle non-conventional artifacts and outperforms the standard approach. Overall, the methodology presented is general and can be applied using various deconvolution tools and in the context of various biological applications, in both human and mouse.

Acknowledgments. This work was supported by the European Research Council (637885) (Y.S., I.G-V), Broad-ISF program (1168/14) (Y.S.) and the Edmond J. Safra Center for Bioinformatics at Tel Aviv University (Y.S.). Research in the I.G-V. lab is supported by the Israeli

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 242–243, 2016.
DOI: 10.1007/978-3-319-31957-5

Science Foundation (1643/13) and the Israeli Centers of Research Excellence (I-CORE): Center No. 41/11. I.G-V. is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and an Alon Fellow.

References

1. Ibrahim, S.F., Van Den Engh, G.: Flow cytometry and cell sorting. In: Advances in Biochemical Engineering/Biotechnology, pp. 19–39 (2007)
2. Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z., Clark, H.F.: Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One **4**, 16 (2009)
3. Altboum, Z., Steuerman, Y., David, E., Barnett-Itzhaki, Z., Valadarsky, L., Keren-Shaul, H., Meningher, T., Mendelson, E., Mandelboim, M., Gat-Viks, I., Amit, I.: Digital cell quantification identifies global immune cell dynamics during influenza infection. Mol. Syst. Biol. **10**, 720 (2014)
4. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., Alizadeh, A.A.: Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods **12**, 453–457 (2015)
5. Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., Butte, A.J.: Cell type–specific gene expression differences in complex tissues. Nat. Methods **7**, 287–289 (2010)

Shall We Dense? Comparing Design Strategies for Time Series Expression Experiments

Emre Sefer and Ziv-Bar Joseph

Department of Computational Biology, School of Computer Science,
Carnegie Mellon University, Pittsburgh, USA
{esefer,zivbj}@cs.cmu.edu

Extended Abstract

Recent advances in sequencing technologies have enabled high throughout profiling of several types of molecular datasets including mRNAs, miRNAs, methylation, and more. Many studies profile one or more of these types of data in a time course. An important experimental design question in such experiments is the number of repeats that is required for accurate reconstruction of the signal being studied. While several studies examined this issue for *static* experiments much less work has focused on the importance of repeats for time series analysis.

Obviously, the more points that can be profiled between the start and end points, the more likely it is that the reconstructed trajectory is accurate. However, in practice the number of time points that are used is usually very small. The main limiting factor is often the budget. While technology has greatly improved, high-throughput NGS studies still cost hundreds of dollars per specific experiment. This is a major issue for time series studies, especially those that need to profile multiple types of biological datasets (mRNA, miRNAs, methylation etc.) at each selected point. Another issue that can limit the number of experiments performed (and so the total number of time points that can be used) is biological sample availability. Thus, when designing such experiments researchers often need to balance the overall goals of reconstructing the most accurate temporal representation of the data types being studied and the need to limit the number of experiments as discussed above.

Given these constraints, an important question when designing high-throughput time-series studies is the need for *repeat* experiments. On the one hand, repeats are a hallmark of biological experiments providing valuable information about noise and reliability of the measured values. On the other, as discussed above, repeats reduce the number of time points that can be profiled which may lead to missing key events between sampled points. Further, if we assume that the biological data being profiled can be represented by a (smooth) continuous curve, which is often the case, then the autocorrelation between successive points can also provide information about noise in the data. In such cases, more time points, even at the expense of fewer or no repeats, may prove to be a better strategy.

Indeed, when looking at datasets deposited in GEO (roughly 25 % of all GEO datasets are time-series), we observe that most of these do not use repeats.

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 244–245, 2016.
DOI: 10.1007/978-3-319-31957-5

However, to the best of our knowledge, no analysis to date was performed to determine the trade-offs between a dense sampling strategy (profiling more time points) and repeat sampling (profiling fewer points, with more than one experiment per point). To study this issue, we use both theoretical analysis and analysis of real data. In our theoretical analysis, we consider a large number of piecewise linear curves and noise levels and compare the expected errors when using the two sampling methods. While the profiles in these biological datasets are usually not piecewise linear, such curves represent important types of biological responses (for example, gradual or single activation, cyclic behavior, increase and then return to baseline, etc.). We also analyze time-series gene expression data to determine the performance of these strategies on real biological data.

Overall, for both, theoretical analysis when using reasonable noise levels and real biological data, we see that dense sampling outperforms repeat sampling indicating that for such data autocorrelation can indeed be a useful feature when trying to reduce the impact of noise on the reconstructed curves. Our results support the commonly used (though so far not justified) practice of reducing or eliminating repeat experiments in time-series high-throughput studies.

Supporting code and datasets: www.cs.cmu.edu/~esefer/genetheoretical

Enabling Privacy Preserving GWAS in Heterogeneous Human Populations

Sean Simmons[1,2], Cenk Sahinalp[2,3], and Bonnie Berger[1]

[1] Department of Mathematics and CSAIL, MIT, Cambridge, MA, USA
bab@mit.edu
[2] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
[3] School of Informatics and Computing, Indiana University, Bloomington, IN, USA

Extended Abstract

With the projected rise of genotyping in the clinic, there has been increasing interest in using patient data to perform genomewide association studies (GWAS) [5, 9]. The idea is to allow doctors and researchers to query patient electronic health records (EHR) to see which diseases are associated with which genomic alterations, avoiding the expense and time required to recruit and genotype patients for a standard GWAS. Such a system, however, leads to major privacy concerns for patients [6]. These privacy concerns have led to tight regulations over who can access this patient data–often it is limited to individuals who have gone through a time consuming application process.

Various approaches have been suggested for overcoming this bottleneck. Specifically, there has been growing interest in using a cryptographic tool known as differential privacy [2] to allow researchers access to this genomic data [3, 4, 8, 11, 12]. Previous approaches for performing differentially private GWAS are based on rather simple statistics that have some major limitations; in particular, they do not correct for a problem known as population stratification, something that is needed when dealing with the genetically diverse populations in many genetic databases. Population stratification is the name given to systematic genomic differences between human populations [10]. It turns out that these differences make it difficult for GWAS to find biologically meaningful associations between common alleles in the population and phenotypes. In order to avoid this problem, various methods have been suggested (EIGENSTRAT [7], LMMs [10], genomic control [1]).

In this work we focus on producing GWAS results that can handle population stratification while still preserving private phenotype information (namely disease status). In particular, we develop a framework that can turn commonly used GWAS statistics (such as LMM based statistics and EIGENSTRAT) into tools for performing privacy preserving GWAS. We demonstrate this framework on one such statistic, EIGENSTRAT [7]. Our method, denoted PrivSTRAT, uses a differentially private framework to protect private phenotype information (disease status) from being leaked while conducting GWAS. Importantly, ours is the first method able to correct for population stratification while preserving privacy in GWAS results. This advance introduces the possibility of applying a

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 246–247, 2016.
DOI: 10.1007/978-3-319-31957-5

differentially private framework to large, genetically diverse groups of patients (such as those present in EHR!).

We test the resulting differentially private EIGENSTRAT statistic, PrivSTRAT, on both simulated and real GWAS datasets to demonstrate its utility. Our results show that for many common GWAS queries, PrivSTRAT is able to achieve high accuracy while enforcing realistic privacy guarantees.

Implementation available at: http://groups.csail.mit.edu/cb/PrivGWAS.

References

1. Devlin, B., Roeder, K.: Genomic control for association studies. Biometrics **55**(4), 997–1004 (1999)
2. Dwork, C., Pottenger, R.: Towards practicing privacy. J. Am. Med. Inform. Assoc. **20**(1), 102–108 (2013)
3. Jiang, X., Zhao, Y., Wang, X., Malin, B., Wang, S., Ohno-Machado, L., Tang, H.: A community assessment of privacy preserving techniques for human genomes. BMC Med. Inform. Decis. Making 14(S1) (2014)
4. Johnson, A., Shmatikov, V.: Privacy-preserving data exploration in genome-wide association studies. In: KDD, pp. 1079–1087 (2013)
5. Lowe, H., Ferris, T., Hernandez, P., Webe, S.: STRIDE - an integrated standards-based translational research informatics platform. In: AMIA Annual Symposium Proceedings, pp. 391–395 (2009)
6. Murphy, S., Gainer, V., Mendis, M., Churchill, S., Kohane, I.: Strategies for maintaining patient privacy in I2B2. JAMIA **18**, 103–108 (2011)
7. Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**, 904–909 (2006)
8. Uhler, C., Fienberg, S., Slavkovic, A.: Privacy-preserving data sharing for genome-wide association studies. J. Priv. Confidentiality **5**(1), 137–166 (2013)
9. Weber, G., Murphy, S., McMurry, A., MacFadden, D., Nigrin, D., Churchill, S., Kohane, I.: The shared health research information network (SHRINE): A prototype federated query tool for clinical data repositories. JAMIA **16**, 624–630 (2009)
10. Yang, J., Zaitlen, N., Goddard, M., Visscher, P., Price, A.: Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. **46**(2), 100–106 (2014)
11. Yu, F., Rybar, M., Uhler, C., Fienberg, S.E.: Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. In: Domingo-Ferrer, J. (ed.) PSD 2014. LNCS, vol. 8744, pp. 170–184. Springer, Heidelberg (2014)
12. Zhao, Y., Wang, X., Jiang, X., Ohno-Machado, L., Tang, H.: Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery. JAMIA **22**, 100–108 (2015)

Efficient Privacy-Preserving Read Mapping Using Locality Sensitive Hashing and Secure Kmer Voting

Victoria Popic(✉) and Serafim Batzoglou

Department of Computer Science, Stanford University, Stanford, CA, USA
{viq,serafim}@stanford.edu

Recent sequencing technology breakthroughs have resulted in an exponential increase in the amount of available sequencing data, enabling major scientific advances in biology and medicine. At the same time, the compute and storage demands associated with processing such datasets have also dramatically increased. Outsourcing computation to commercial low-cost clouds provides a convenient and cost-effective solution to this problem. However, exposing genomic data to an untrusted third-party also raises serious privacy concerns [1]. Read alignment is a critical and computationally intensive first step of most genomic data analysis pipelines. While significant effort has been dedicated to optimize this task, few approaches have addressed outsourcing this computation securely to an untrusted party. The few secure solutions that exist either do not scale to whole genome sequencing datasets [2] or are not competitive with the state of the art in read mapping [3].

In this work we present BALAUR, a privacy preserving read mapping technique that securely outsources a significant portion of the read-mapping task to the public cloud, while being highly competitive with existing state-of-the-art aligners. Our approach is to reduce the alignment task to a secure voting procedure based on matches between read and reference kmers, taking advantage of the high similarity between the reads and their corresponding positions in the reference. At a high level, BALAUR can be summarized in the following two phases: (1) fast identification of a few candidate alignment positions in the genome using the locality sensitive hashing scheme MinHash [4] on the private client and (2) secure kmer voting against each such candidate position to determine the optimal read mappings on the public server. To outsource Phase 2 securely to the cloud, voting is performed using encrypted kmers of each read and its selected reference candidate contigs. In order to prevent frequency attacks using background knowledge (e.g. kmer repeat statistics), our encryption scheme uses the traditional cryptographic hashing scheme SHA-1, along with unique per-read keys and intra-read repeat masking, which prevents the adversary from detecting kmers that are equal across and inside each read or contig. We compare the performance of BALAUR with several popular and efficient non-cryptographic state-of-the-art read aligners, such as BWA-MEM [5] and Bowtie 2 [6], using simulated and real whole human genome sequencing datasets. We demonstrate that our approach achieves

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 248–249, 2016.
DOI: 10.1007/978-3-319-31957-5

similar accuracy and runtime performance on typical short-read datasets, while being significantly faster than state of the art in long read mapping.

References

1. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. **15**(6), 409–421 (2014)
2. Huang, Y., Evans, D., Katz, J., Malka, L.: Faster secure two-party computation using garbled circuits. In: USENIX Security Symposium, vol. 201 (2011)
3. Chen, Y., Peng, B., Wang, X., Tang, H.: Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. In: NDSS (2012)
4. Broder, A.Z., Charikar, M., Frieze, A.M., Mitzenmacher, M.: Min-wise independent permutations. J. Comput. Syst. Sci. **60**(3), 630–659 (2000)
5. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 (2013)
6. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**(4), 357–359 (2012)

Finding Mutated Subnetworks Associated with Survival in Cancer

Tommy Hansen[1] and Fabio Vandin[1,2,3]

[1] Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
tvhansen33@gmail.com, vandinfa@dei.unipd.it
[2] Department of Information Engineering, University of Padova, Padova, Italy
[3] Department of Computer Science, Brown University, Providence, RI, USA

Next-generation sequencing technologies allow the measurement of somatic mutations in a large number of patients from the same cancer type. One of the main goals in the analysis of these mutations is the identification of mutations associated with clinical parameters, for example survival time. This goal is hindered by the extensive genetic heterogeneity in cancer, with different genes mutated in different patients. This heterogeneity is due to the fact that genes and mutations act in the context of *pathways*, and it is therefore crucial to study mutations in the context of interactions among genes. In this work we study the problem of identifying subnetworks of a large gene-gene interaction network that have somatic mutations associated with survival from genome-wide mutation data of a large cohort of cancer patients. We formally define the associated computational problem by using a score for subnetworks based on the test statistic of the log-rank test, a widely used statistical test for comparing the survival of two given populations. We show that the computational problem is NP-hard in general and that it remains NP-hard even when restricted to graphs with at least one node of large degree, the case of interest for gene-gene interaction networks.

We propose a novel randomized algorithm, called Network of Mutations Associated with Survival (NoMAS), to find subnetworks of a large interaction network whose mutations are associated with survival time. NoMAS is based on the color-coding technique, but differently from previous applications of color-coding our score is not additive, therefore NoMAS does not inherit the guarantees given by color-coding for the identification of the optimal solution. Nonetheless, we prove that under a reasonable model for mutations in cancer NoMAS does identify the optimal solution with high probability when the subnetwork size is not too large and given mutations from a sufficiently large number of patients. We implemented NoMAS and tested it on simulated and cancer data. The results show that our method does indeed find the optimal solution and performs better than greedy approaches commonly used to solve optimization problems on networks. Moreover, on two large cancer datasets NoMAS identifies subnetworks with significant association to survival, while none of the genes in the subnetwork has significant association with survival when considered in isolation.

This work is supported, in part, by the University of Padova under project CPDA121378/12 and by NSF grant IIS-1247581.

M. Singh (Ed.): RECOMB 2016, LNBI 9649, p. 250, 2016.
DOI: 10.1007/978-3-319-31957-5

Multi-state Perfect Phylogeny Mixture Deconvolution and Applications to Cancer Sequencing

Mohammed El-Kebir[1], Gryte Satas[1], Layla Oesper[1,2], and Benjamin J. Raphael[1]

[1] Center for Computational Molecular Biology and Department of Computer Science, Brown University, Providence, RI 02912, USA
[2] Department of Computer Science, Carleton College, Northfield, MN 55057, USA

Abstract. The reconstruction of phylogenetic trees from mixed populations has become important in the study of cancer evolution, as sequencing is often performed on bulk tumor tissue containing mixed populations of cells. Recent work has shown how to reconstruct a perfect phylogeny tree from samples that contain mixtures of two-state characters, where each character/locus is either mutated or not. However, most cancers contain more complex mutations, such as copy-number aberrations, that exhibit more than two states. We formulate the Multi-State Perfect Phylogeny Mixture Deconvolution Problem of reconstructing a multi-state perfect phylogeny tree given mixtures of the leaves of the tree. We characterize the solutions of this problem as a restricted class of spanning trees in a graph constructed from the input data, and prove that the problem is NP-complete. We derive an algorithm to enumerate such trees in the important special case of cladisitic characters where the ordering of the states of each character is given. We apply our algorithm to simulated data and to two cancer datasets. On simulated data, we find that for a small number of samples, the Multi-State Perfect Phylogeny Mixture Deconvolution Problem often has many solutions, but that this ambiguity declines quickly as the number of samples increases. On real data, we recover copy-neutral loss of heterozygosity, single-copy amplification and single-copy deletion events, as well as their interactions with single-nucleotide variants.

M. Singh (Ed.): RECOMB 2016, LNBI 9649, p. 251, 2016.
DOI: 10.1007/978-3-319-31957-5

Tree Inference for Single-Cell Data

Katharina Jahn[1,2], Jack Kuipers[1,2], and Niko Beerenwinkel[1,2]

[1] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2] SIB Swiss Institute of Bioinformatics, Basel, Switzerland

The genetic heterogeneity found within tumour cells is considered a major cause for the development of drug resistance during cancer treatment. Subclonal cell populations may possess a distinct set of genetic lesions that render them non-susceptible to the selected therapy resulting in an eventual tumour regrowth originating from the surviving cells. To develop more efficient therapies it is therefore paramount to understand the subclonal structure of the individual tumour along with its mutational history.

Classical next-generation sequencing techniques provide admixed mutation profiles of millions of cells whose deconvolution into subclones is often an underdetermined problem that limits the resolution at which the subclonal composition can be reconstructed. Recent technological advances now allow for the sequencing of individual cells. While this progress comes at the cost of higher error rates, it still provides the possibility to reconstruct mutational tumour histories at an unprecedented resolution.

We present a stochastic search algorithm to identify the evolutionary history of a tumour from noisy and incomplete mutation profiles of single cells. Our approach, termed SCITE, comprises a flexible MCMC sampling scheme that allows us to compute the maximum likelihood mutation tree and to sample from its posterior probability distribution. Tree reconstruction can include attachment of the single-cell samples and can be combined with estimating the error rates of the sequencing experiments. We evaluate SCITE on real cancer data showing its scalability to present day single-cell sequencing data and improved accuracy in tree reconstruction over existing approaches. In addition, we estimate from simulation studies the number of cells necessary for reliable mutation tree reconstruction which could inform the design of future single-cell sequencing projects.

K. Jahn and J. Kuipers—Equal contributors

M. Singh (Ed.): RECOMB 2016, LNBI 9649, p. 252, 2016.
DOI: 10.1007/978-3-319-31957-5

mLDM: A New Hierarchical Bayesian Statistical Model for Sparse Microbial Association Discovery

Yuqing Yang[1,2], Ning Chen[1], and Ting Chen[1,2,3,4]

[1] Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST, Beijing, China
{ningchen,tingchen}@tsinghua.edu.cn
[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3] State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China
[4] Program in Computational Biology and Bioinfomatics, University of Southern California, Los Angeles, CA 90089, USA

Understanding associations among microbes and associations between microbes and their environmental factors from metagenomic sequencing data is a key research topic in microbial ecology, which could help us to unravel real interactions (e.g., commensalism, parasitism, competition, etc.) in a community as well as understanding community-wide dynamics. Although several statistical tools have been developed for metagenomic association studies, they either suffer from compositional bias or fail to take into account environmental factors that directly affect the composition of a microbial community, leading to some false positive associations. For example, two unrelated microbes may appear to be associated just because they both respond to the same environmental perturbation.

We propose metagenomic Lognormal-Dirichlet-Multinomial (mLDM), a hierarchical Bayesian model with sparsity constraints to bypass compositional bias and discover new associations among microbes and associations between microbes and their environmental factors. mLDM is able to: (1) infer both conditionally dependent associations among microbes and direct associations between microbes and environmental factors; (2) consider both compositional bias and variance of metagenomic data; and (3) estimate absolute abundance for microbes. These associations can capture the direct relationships underlying pairs of microbes and remove the indirect connections induced from other common factors. mLDM discovers the metagenomic associations using a hierarchical Bayesian graphical model with sparse constraints, where the metagenomic sequencing data generating process is captured by the hierarchical latent variable model. Specifically, we assume that the read counts are proportional to

This paper was selected for oral presentation at RECOMB 2016 and an abstract is published in the conference proceedings. The work is supported by the NSFC grant (Nos: 61305066, 61561146396, 61332007, 61322308), NIH grant (NIH/NHGRI 1U01 HG006531-01) and NSF grants (NSF/OCE 1136818 and NSF/DMS ATD 7031026).

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 253–254, 2016.
DOI: 10.1007/978-3-319-31957-5

the latent microbial ratios which are determined by their absolute abundance. The microbial absolute abundance is influenced by two factors: (1) environmental factors, whose effects on the microbes are denoted by a linear regression model; and (2) the associations among microbes encoded by a latent vector, which is determined by the matrix that records microbial associations and the mean vector that affects the basic absolute abundance of microbes. By introducing sparsity regularization, mLDM can capture both the conditionally dependent associations among microbes and the direct associations between microbes and environmental factors. The task is formulated as solving an optimization problem, which can be solved using coordinate descent or proximal methods. For model selection, we choose the best parameters via extended Bayesian information criteria (EBIC).

To show the effectiveness of the proposed mLDM model, we conducted several experiments using synthetic data, the western English Channel time-series sequencing data, and the Ocean TARA data, and compared it with several state-of-the-art methodologies, including PCC, SCC, LSA, CCREPE, SparCC, CCLasso, glasso (graphical lasso), SPIEC-EASI (mlasso) and SPIEC-EASI (glasso). The results demonstrate that the association network computed by the mLDM model, is closest to the true network, and that the mLDM model can recover most of the conditionally dependent associations. For the latter two experimental datasets, mLDM can discover most known interactions in addition to several potentially interesting associations.

Low-Density Locality-Sensitive Hashing Boosts Metagenomic Binning

Yunan Luo[1,4], Jianyang Zeng[1], Bonnie Berger[2,3], and Jian Peng[4]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
bab@mit.edu
[3] Department of Mathematics, MIT, Cambridge, MA, USA
[4] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
jianpeng@illinois.edu

1 Introduction

Metagenomic sequencing techniques produce large data sets of DNA fragments (e.g. reads or contigs) from environmental samples. To understand the microbial communities and functional structures within the samples, metagenomic sequence fragments need to be first assigned to their taxonomic origins from which they were derived (also called "binning") to facilitate downstream analyses.

Arguably the most popular metagenomic binning approaches are alignment-based methods. A sequence fragment is searched against a reference database consisting of full genomes of organisms, and the highest scoring organism is assigned as the taxonomic origin. Although efficient sequence alignment algorithms, including BWA-MEM [1], Bowtie2 [2] and (mega)BLAST [3], can readily be used for this purpose, the computational cost of alignment-based methods becomes prohibitive as the size of the sequence dataset grows dramatically, which is often the case in recent studies.

Another completely different binning approach is based on genomic sequence composition, which exploits the sequence characteristics of metagenomic fragments and applies machine learning classification algorithms to assign putative taxnomic origins to all fragments. Since classifiers, such as support vector machines, are trained on whole reference genome sequences beforehand, compositional methods normally are substantially faster than alignment-based methods on large datasets. The rationale behind compositional-based binning methods is based on the fact that different genomes have different conserved sequence composition patterns, such as GC content, codon usage or a particular abundance distribution of consecutive nucleotide k-mers. To design a good compositional-based algorithm, we need to extract informative and discriminative features from the reference genomes. Most existing methods, including PhyloPythia(S) [4, 5], use k-mer frequencies to represent sequence fragments, where k is typically small (e.g. 6 to 10). While longer k-mers, which capture compositional dependency

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 255–257, 2016.
DOI: 10.1007/978-3-319-31957-5

within larger contexts, could potentially lead to higher binning accuracy, they are more prone to noise and errors if used in the supervised setting. Moreover, incorporating long k-mers as features increases computational cost exponentially and requires significantly larger training datasets.

2 Method

We introduce a novel compositional metagenomic binning algorithm, Opal, which robustly represents long k-mers in a compact way to better capture the long-range compositional dependencies in a fragment. The key idea behind our algorithm is built on locality-sensitive hashing (LSH), a dimensionality-reduction technique that hashes input high-dimensional data into low-dimensional buckets, with the goal of maximiz the probability of collisions for similar input data. To the best of our knowledge, it is the first time that LSH functions have been applied for compositional-based metagenomic binning. We propose to use them first to represent metagenomic fragments compactly and subsequently for machine learning classification algorithms to train metagenomic binning models. Since metagenomic fragments can be very long, sometimes from hundreds of bps to tens of thousands of bps, we hope to construct compositional profiles to encode long-range dependencies within long k-mers. To handle large ks, we develop string LSH functions to compactly encode global dependencies with k-mers in a low-dimensional feature vector, as oppose to directly using a 4^k-length k-mer profile vector. Although LSH functions are usually constructed in a uniformly random way, we propose a new and efficient design of LSH functions based on the idea of the low-density parity-check (LDPC) code invented by Robert G Gallager for noisy message transmission [6, 7]. A key observation is that Gallager's LDPC design not only leads to a family of LSH functions but also makes them efficient such that even a small number of random LSH functions can effectively encode long fragments. Different from uniformly random LSH functions, the Gallager LSH functions are constructed structurally and hierarchically to ensure the compactness of the feature representation and robustness when sequencing noise appears in the data. Methodologically, starting from a Gallager design matrix with row weight t, we construct m hash functions to encode high-order sequence compositions within a k-mer. In contrast to the $O(4^k)$ complexity it would take to represent contiguous k-mers, our proposed Gallager LSH adaptation requires only $O(m4^t)$ time. For very long k-mers, we construct the Gallager LSH functions in a hierarchical fashion to further capture compositional dependencies from both local and global contexts. It is also possible to use Opal as a "coarse search" procedure in the compressive genomics manner to reduce the search space of alignment-based methods [8]. We first apply the compositional-based binning classifier to identify a very small subset or group of putative taxonomic origins which are ranked very highly by the classifier. Then we perform sequence alignment between the fragment and the reference genomes of the top-ranked organisms. This natural combination of compositional-based and alignment-based methods provides metagenomic binning with high scalability, high accuracy and high-resolution alignments.

3 Results

To evaluate the performance of Opal, we trained an SVM model with features generated by the Gallager LSH method. When tested on a large dataset of 50 microbial species, Opal achieved better binning accuracy than the traditional method that uses contiguous k-mer profiles as features [4]. Moreover, our method is more robust to mutations and sequencing errors, compared to the method with the contiguous k-mer representation. Opal outperformed (in terms of robustness and accuracy) BWA-MEM [1], the state-of-the-art alignment-based method. Remarkably, we achieved up to two orders of magnitude improvement in binning speed on large datasets with mutations rates ranging from 5 % to 15 % over 20–50 microbial species; moreover, we found Opal to be substantially more accurate than BWA-MEM when the rate of sequencing error is high (e.g., 10–15 %). It is counterintuitive that a compositional binning approach is as robust as or even more robust than alignment-based approaches, particularly in the presence of high sequencing errors or mutations in metagenomic sequence data. Finally, we combined both compositional and alignment-based methods, by applying the compositional SVM with the Gallager LSH coding as a "coarse-search" procedure to reduce the taxonomic space for a subsequent alignment-based BWA-MEM "fine search." This integrated approach is almost 20 times faster than original BWA-MEM and also has substantially improved binning performance on noisy data. The above results indicate that Opal enables us to perform accurate metagenomic analysis for very large metagenomic studies with greatly reduced computational cost.

Acknowledgments. This work was partially supported by the US National Institute of Health Grant GM108348, the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003 and 61472205.

References

1. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv preprints (2013). arXiv:1303.3997
2. Langmead, B., Salzberg, S.: Fast gapped-read alignment with bowtie 2. Nat. Methods **9**, 357–359 (2012)
3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**, 403–410 (1990)
4. Patil, K.R., Haider, P., Pope, P.B., Turnbaugh, P.J., Morrison, M., Scheffer, T., McHardy, A.C.: Taxonomic metagenome sequence assignment with structured output models. Nat. Methods **8**(3), 191–192 (2011)
5. McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I.: Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods **4**(1), 63–72 (2007)
6. Gallager, R.: Low-density parity-check codes. IEEE Trans. Inf. Theory **8**(1), 21–28 (1962)
7. MacKay, D., Neal, R.: Near shannon limit performance of low density parity check codes. Electron. Lett. **32**, 1645–1646 (1996)
8. Yu, Y.W., Daniels, N.M., Danko, D.C., Berger, B.: Entropy-scaling search of massive biological data. Cell Syst. **2**, 130–140 (2015)

metaSPAdes: A New Versatile *de novo* Metagenomics Assembler

Sergey Nurk[1], Dmitry Meleshko[1], Anton Korobeynikov[1], and Pavel Pevzner[1,2]

[1] Center for Algorithmic Biotechnology, Institute for Translational Biomedicine, Saint Petersburg State University, Saint Petersburg, Russia
[2] Department of Computer Science and Engineering, University of California, San Diego, USA
ppevzner@ucsd.edu

Metagenome sequencing has emerged as a technology of choice for analyzing bacterial populations and discovery of novel organisms and genes. While many metagenomics assemblers have been developed recently, assembly of metagenomic data remains difficult thus stifling biological discoveries.

We developed metaSPAdes tool that addresses the specific challenges of metagenomic assembly by combining new algorithmic ideas with methods previously proposed for assembling single cells [1] and highly polymorphic genomes [2].

metaSPAdes features (i) efficient analysis of strain mixtures, (ii) a novel repeat resolution approach that utilizes the local read coverage of the regions that are being reconstructed, (iii) a novel algorithm that, somewhat counterintuitively, utilizes strain differences to improve reconstruction of the consensus genomes of a strain mixture, and (iv) improved running time and reduced memory footprint to enable assemblies of large metagenomes.

We benchmarked metaSPAdes against the state-of-the-art metagenomics assemblers (MEGAHIT [3], IDBA-UD [4] and Ray-Meta [5]) across diverse datasets and demonstrated that it results in high-quality assemblies.

References

1. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A., Dvorkin, M., Kulikov, A., Lesin, V., Nikolenko, S., Pham, S., Prjibelski, A., Pyshkin, A., Sirotkin, A., Vyahhi, N., Tesler, G., Alekseyev, M., Pevzner, P.: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. **19**(5), 455–477 (2012)
2. Safonova, Y., Bankevich, A., Pevzner, P.: dipSPAdes: assembler for highly polymorphic diploid genomes. J. Comput. Biol. **22**(6), 528–545 (2015)
3. Li, D., Liu, C.M., Luo, R., Sadakane, K., Lam, T.W.: MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics **31**(10), 1674–1676 (2015)
4. Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L.: IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28**(11), 1420–1428 (2012)
5. Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., Corbeil, J.: Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. **13**(12), R122 (2012)

M. Singh (Ed.): RECOMB 2016, LNBI 9649, p. 258, 2016.
DOI: 10.1007/978-3-319-31957-5

Distributed Gradient Descent in Bacterial Food Search

Shashank Singh[1], Sabrina Rashid[2], Saket Navlakha[3], and Ziv Bar-Joseph[4]

[1] Machine Learning Department and Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2] Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] Integrative Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[4] Machine Learning Department and Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
zivbj@cs.cmu.edu

Extended Abstract

Communication and coordination play a major role in the ability of bacterial cells to adapt to changing environments and conditions. Recent work has shown that such coordination underlies several aspects of bacterial responses including their ability to develop antibiotic resistance. Here we show that a variant of a commonly used machine learning algorithm, *distributed gradient descent*, is utilized by large bacterial swarms to efficiently search for food when faced with obstacles in their environment. Similar to conventional gradient descent, by sensing the food gradient, each cell has its own belief about the location of the food source. However, given limits on the ability of each cell to accurately detect and move toward the food source in a noisy environment with obstacles, the individual trajectories may not produce the optimal path to the food source. Thus, in addition to using their own belief each cell also sends and receives messages from other cells (either by secreting specific proteins or by physical interaction), which are integrated to update its belief and determine its direction and velocity. The process continues until the swarm converges to the food source.

Our main contribution is to better understand the computation performed by cells during collective foraging. Current models of this process are largely based on differential equation methods which do not fully take into account how the topology of the cellular interaction network changes over time. Furthermore, the assumptions made by these models about the ability of cells to identify the source(s) of the messages and to utilize a large (effectively continuous valued) set of messages, are unrealistic given the limited computational powers bacteria cells possess.

Here, we develop a distributed gradient descent algorithm that makes biologically realistic assumptions regarding the dynamics of the cells, the size of the

S. Singh and S. Rashid—These authors contributed equally.

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 259–260, 2016.
DOI: 10.1007/978-3-319-31957-5

messages communicated, and their ability to identify senders, while still solving the bacterial food search problem more efficiently (in terms of the overall complexity of messages sent) and more quickly (in terms of the time it takes the swarm to reach the food source) when compared to current differential equation models. We prove that our model converges to a local minimum, under reasonable assumptions on how bacteria communicate and perform simulation studies and analysis of experimental data. These experiments indicate that our communication model is feasible and leads to improvements over prior methods and over single cell and single swarm behavior.

There are many parallel requirements of computational and biological systems, suggesting that each can learn from the other. We conclude by discussing how the efficient and robust bacterial gradient descent algorithms we developed can be used by distributed sensors or wireless networks that operate under strict communication and computation constraints.

Supporting movies: www.andrew.cmu.edu/user/sabrinar/Bacteria_Simulation_Movies/

AptaTRACE: Elucidating Sequence-Structure Binding Motifs by Uncovering Selection Trends in HT-SELEX Experiments

Phuong Dao[1], Jan Hoinka[1], Yijie Wang[1], Mayumi Takahashi[2], Jiehua Zhou[2], Fabrizio Costa[3], John Rossi[2], John Burnett[2], Rolf Backofen[3], and Teresa M. Przytycka[1(✉)]

[1] National Center of Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894 , USA
przytyck@ncbi.nlm.nih.gov
[2] Department of Molecular and Cellular Biology, Beckman Research Institute of City of Hope, Duarte, CA, USA
[3] Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany

Abstract. Aptamers, short synthetic RNA/DNA molecules binding specific targets with high affinity and specificity, are utilized in an increasing spectrum of bio-medical applications. Aptamers are identified *in vitro* via the Systematic Evolution of Ligands by Exponential Enrichment (SELEX) protocol. SELEX selects binders through an iterative process that, starting from a pool of random ssDNA/RNA sequences, amplifies target-affine species through a series of selection cycles. HT-SELEX, which combines SELEX with high throughput sequencing, has recently transformed aptamer development and has opened the field to even more applications. HT-SELEX is capable of generating over half a billion data points, challenging computational scientists with the task of identifying aptamer properties such as sequence-structure motifs that determine binding. While currently available motif finding approaches suggest partial solutions to this question, none possess the generality or scalability required for HT-SELEX data, and they do not take advantage of important properties of the experimental procedure.

We present AptaTRACE, a novel approach for the identification of sequence-structure binding motifs in HT-SELEX derived aptamers. Our approach leverages the experimental design of the SELEX protocol and identifies sequence-structure motifs that show a signature of selection towards a preferred structure. In the initial pool, secondary structural contexts of each k-mer are distributed according to a background distribution. However, for sequence motifs involved in binding, in later selection cycles, this distribution becomes biased towards the structural context favored by the binding interaction with the target site. Thus, AptaTRACE aims at identifying sequence motifs whose tendency of residing in a hairpin, bugle loop, inner loop, multiple loop, dangling end, or of being paired converges to a specific structural context throughout the selection cycles

P. Dao and J. Hoinka—Equal contribution, these authors are listed in alphabetical order.

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 261–262, 2016.
DOI: 10.1007/978-3-319-31957-5

of HT-SELEX experiments. For each k-mer, we compute the distribution of its structural contexts in each sequenced pool. Then, we compute the relative entropy (KL-divergence) based score, to capture the change in the distribution of its secondary structure contexts from a cycle to a later cycle. The relative entropy based score is thus an estimate of the selection towards the preferred secondary structure(s).

We show our results of applying AptaTRACE to simulated data and an *in vitro* selection consisting of high-throughput data from 9 rounds of cell-SELEX. In testing on simulated data, AptaTRACE outperformed other generic motif finding methods in terms of sensitivity. By measuring selection towards sequence-structure motifs by the change in their distributions of the structural contexts and not based on abundance, AptaTRACE can uncover motifs even when these are present only in a small fraction of the pool. Moreover, our method can also help to reduce the number of selection cycles required to produce aptamers with the desired properties, thus reducing cost and time of this rather expensive procedure.

Fast Bayesian Inference of Copy Number Variants Using Hidden Markov Models with Wavelet Compression

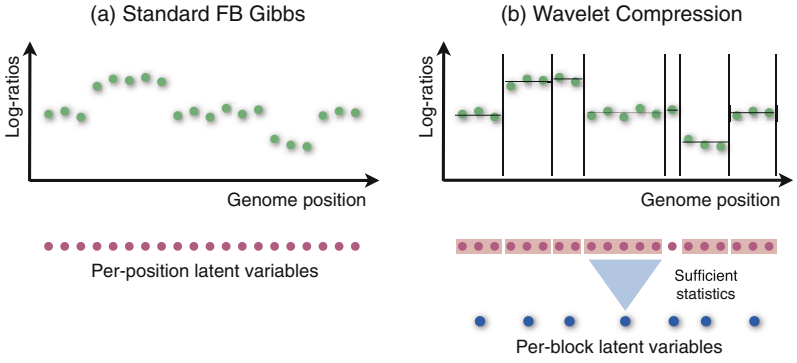
John Wiedenhoeft, Eric Brugel, and Alexander Schliep

Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA
{john.wiedenhoeft,alexander}@schlieplab.org

Hidden Markov Models (HMM) are statistical models frequently used in Copy Number Variant (CNV) detection. Classic frequentist maximum likelihood techniques for parameter estimation like Baum-Welch are not guaranteed to be globally optimal, and state inference via the Viterbi algorithm only yields a single MAP segmentation. Nevertheless, Bayesian methods like Forward-Backward Gibbs sampling (FBG) are rarely used due to long running times and slow convergence.

Here, we exploit that both state sequence inference and wavelet regression reconstruct a piecewise constant function from noisy data, though under different constraints. We draw upon a classic minimaxity result from wavelet theory to dynamically compress the data into segments of successive observation whose variance can be explained as emission noise under the current parameters in each FBG iteration, and are thus unlikely to yield state transitions indicating a break point. We further show that such a compression can be rapidly recomputed with little overhead using a simple data structure. Due to the summary treatment of subsequent observations in segments (or blocks) derived from the wavelet regression—see panels (a) and (b) below—we simultaneously achieve drastically reduced running times as well as improved convergence behavior of FBG. To the best of our knowledge this shows for the first time that a fully Bayesian HMM can be competitive with or outperform the current state of the art.

This makes routine diagnostic use and re-analysis of legacy data collections feasible; to this end, we also propose an effective automatic prior. An open source software implementation is available at http://schlieplab.org/Software/HaMMLET/.



M. Singh (Ed.): RECOMB 2016, LNBI 9649, p. 263, 2016.
DOI: 10.1007/978-3-319-31957-5

Allele-Specific Quantification of Structural Variations in Cancer Genomes

Yang Li[1], Shiguo Zhou[2], David C. Schwartz[2], and Jian Ma[1,3,4]

[1] Department of Bioengineering, University of Illinois at Urbana-Champaign, Champaign, USA
[2] Laboratory for Molecular and Computational Genomics, University of Wisconsin-Madison, Madison, USA
[3] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, USA
[4] School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
jianma@cs.cmu.edu

One of the hallmarks of cancer genome is aneuploidy, which causes abnormal copy numbers of alleles. Structural variations (SVs) can further modify the aneuploid cancer genomes into a mixture of rearranged genomic segments with extensive range of somatic copy number alterations (CNAs). Indeed, aneuploid cancer genomes have significantly higher rate of CNAs and SVs. However, although methods have been developed to identify SVs and allele-specific copy number of genome (ASCNG) separately, no existing algorithm can simultaneously analyze SVs and ASCNG. Such integrated approach is particularly important to fully understand the complexity of cancer genomes.

In this work, we introduce a novel computational method `Weaver` to identify allele-specific copy number of SVs (ASCNS) as well as the inter-connectivity of them in aneuploid cancer genomes. To our knowledge, this is the first method that can simultaneously analyze SVs and ASCNG. Under the same method framework, `Weaver` also provides base-pair resolution ASCNG. Note that in this work we specifically focus on the quantification of SV copy numbers, which is the key novelty of our method. Our framework is flexible to allow users to choose their own variant calling (including SV) tools. We use the variant calling results to build a cancer genome graph, which is subsequently converted to a pairwise Markov Random Field (MRF). In the MRF, the ASCNS and SV phasing configuration, together with ASCNG, are hidden states in the nodes and the observations contain all sequencing information, including coverage, read linkage between SNPs as well as connections between SVs and SNPs. Therefore, our goal of finding the ASCNS and SV phasing together with ASCNG is formulated as searching the *maximum a posteriori* (MAP) solution for MRF. We apply Loopy Belief Propagation (LBP) framework to solve the problem.

We extensively evaluated the performance of `Weaver` using simulation. We also compared with single-molecule Optical Mapping analysis and evaluated using real data (including MCF-7, HeLa, and TCGA whole genome sequencing samples). We demonstrated that `Weaver` is highly accurate and can greatly refine the analysis of complex cancer genome structure. We believe `Weaver` provides a more integrative solution to study complex cancer genomic alterations.

M. Singh (Ed.): RECOMB 2016, LNBI 9649, p. 264, 2016.
DOI: 10.1007/978-3-319-31957-5

Assembly of Long Error-Prone Reads Using de Bruijn Graphs

Yu Lin[1], Max W. Shen[1], Jeffrey Yuan[1],
Mark Chaisson[2], and Pavel A. Pevzner[1]

[1] Department of Computer Science and Engineering,
University of California San Diego, San Diego, USA
[2] Department of Genome Sciences, University of Washington,
Washington, D.C., USA

When the first reads generated using Single Molecule Real Time (SMRT) technology were made available, most researchers were skeptical about the ability of existing algorithms to generate high-quality assemblies from error-prone SMRT reads. Roberts et al. [3] even referred to this widespread skepticism as the error myth and argued that new assemblers for error-prone reads need to be developed to debunk this myth.

Recent algorithmic advances resulted in accurate assemblies from error-prone reads generated by Pacific Biosciences and even from less accurate Oxford Nanopore reads. However, previous studies of SMRT assemblies were based on the overlap-layout-consensus (OLC) approach, which dominated genome assembly in the last decade, is inapplicable to assembling long reads. This is a misunderstanding since the de Bruijn approach, as well as its variation called the *A-Bruijn* graph approach [2], was originally developed to assemble rather long Sanger reads.

There is also a misunderstanding that the de Bruijn graph approach can only assemble highly accurate reads and fails while assembling error-prone SMRT reads, yet another error myth that we debunk. The A-Bruijn graph approach was originally designed to assemble inaccurate reads as long as any similarities between reads can be reliably identified. However, while A-Bruijn graphs have proven to be useful in assembling Sanger reads and mass spectra (highly inaccurate fingerprints of amino acid sequences of peptides [1]), the question of how to apply A-Bruijn graphs for assembling SMRT reads remains open. We show how to generalize de Bruijn graphs to assemble long error-prone reads and describe the ABruijn assembler, which results in more accurate genome reconstructions than the state-of-the-art algorithms for assembling Pacific Biosciences and Oxford Nanopore reads.

References

1. Bandeira, N., Pham, V., Pevzner, P., Arnott, D., Lill, J.R.: Automated de novo protein sequencing of monoclonal antibodies. Nat. Biotechnol. **26**, 1336–1338 (2008)
2. Pevzner, P.A., Tang, H., Tesler, G.: De novo repeat classification and fragment assembly. Genome Res. **14**, 1786–1796 (2004)
3. Roberts, R.J., Carneiro, M.O., Schatz, M.C.: The advantages of SMRT sequencing. Genome Biol. **14**, 405 (2013)

M. Singh (Ed.): RECOMB 2016, LNBI 9649, p. 265, 2016.
DOI: 10.1007/978-3-319-31957-5

Locating a Tree in a Reticulation-Visible Network in Cubic Time

Andreas D.M. Gunawan[1], Bhaskar DasGupta[2], and Louxin Zhang[1]

[1] Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
matzlx@nus.edu.sg

[2] Department of Mathematics, National University of Singapore, Singapore 119076, Singapore

In studies of molecular evolution, phylogenetic trees are rooted trees, whereas phylogenetic networks are rooted acyclic digraphs. Edges are directed away from the root and leaves are uniquely labeled with taxa in phylogenetic networks. An important bioinformatics task is checking the "consistency" of two evolutionary models. This has motivated researchers to study the problem of determining whether a tree is displayed by a network or not, which is called the tree containment problem (TCP) [2, 3]. The cluster containment problem (CCP) is related algorithmic problem that asks whether or not a subset of taxa is a cluster in a tree displayed by a network [2].

Both the TCP and CCP are NP-complete [3], even on a very restricted class of networks [4]. An open questions was posed by van Iersel *et al.* asking whether or not the TCP is solvable in polynomial time for binary reticulation-visible networks [1, 2, 4]. A network is reticulation-visible if every reticulation separates the root of the network from some leaves [2], where reticulations are internal nodes of indegree greater than one and outdegree one.

We give an affirmative answer to the open problem of van Iersel, Semple and Steel by presenting a cubic time algorithm for the TCP for arbitrary reticulation-visible networks. The key tool used in our answer is a powerful decomposition theorem. It also allows us to design a linear-time algorithm for the cluster containment problem for networks of this type and to prove that every galled network with n leaves has $2(n-1)$ reticulation nodes at most. The full version of this work can be found at arXiv.org (arXiv:1507.02119v2).

References

1. Gambette, P., Gunawan, A.D.M., Labarre, A., Vialette, S., Zhang, L.: Locating a tree in a phylogenetic network in quadratic time. In: Przytycka, T.M. (ed.) RECOMB 2015. LNCS, vol. 9029, pp. 96–107. Springer, Heidelberg (2015)
2. Huson, D.H., Rupp, R., Scornavacca, C.: Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge University Press, Cambridge (2011)
3. Kanj, I.A., Nakhleh, L., Than, C., Xia, G.: Seeing the trees and their branches in the network is hard. Theor. Comput. Sci. **401**, 153–164 (2008)
4. van Iersel, L., Semple, C., Steel, M.: Locating a tree in a phylogenetic network. Inform. Process. Lett. **110**, 1037–1043 (2010)

M. Singh (Ed.): RECOMB 2016, LNBI 9649, p. 266, 2016.
DOI: 10.1007/978-3-319-31957-5

Joint Alignment of Multiple Protein-Protein Interaction Networks via Convex Optimization

Somaye Hashemifar, Qixing Huang, and Jinbo Xu

Toyota Technological Institute at Chicago, Chicago, USA
{somaye.hashemifar,pqx.huang,jinboxu}@gmail.com

Abstract. Protein-protein interaction (PPI) network alignment greatly benefits the understanding of evolutionary relationships among species and identifying conserved sub-networks. Although a few methods have been developed for multiple PPI networks alignment, the alignment quality is still far away from perfect. This paper presents a new method ConvexAlign for joint alignment of multiple PPI networks that can generate functionally much more consistent alignments than existing methods.

1 Introduction

This paper presents a novel method ConvexAlign for one-to-one global network alignment (GNA). A one-to-one alignment is a mapping in which one protein is cannot be aligned more than one protein in another network. ConvexAlign calculates the optimal alignment by maximizing a scoring function that integrates sequence similarity, network topology and interaction preserving. We formulate the problem as an integer program and relax it to a convex optimization problem, which enables us to simultaneously align all the PPI networks, without resorting to the widely-used seed-and-extension or progressive alignment methods. Then we use ADMM to solve the relaxed convex optimization problem. Our results show that ConvexAlign outperforms several popular alignment methods both topologically and biologically.

2 Method

Scoring function. Let $G = (V, E)$ denote a PPI network where V is the set of vertices (proteins) and E is the set of edges (interactions). A one-to-one multiple alignment between N networks is given by a binary matrix X where $X_{ij}(v_i, v_j) = 1$ if and only if v_i and v_j are aligned and there is at most one 1 in each row or column of X. It is easy to see X is positive semi-definite. Let C represent a matrix where each value C_{ij} indicates the similarity between two proteins v_i and v_j. Our goal is to find an alignment that maximizes the number of matched orthologous proteins and the number of preserved edges. We define the node score of an alignment $\mathcal{A}$ as follows: $f_{node}(\mathcal{A}) = \sum_{1 \le i < j \le N} C_{ij}, X_{ij}$.
We also define the edge score to count the number of preserved edges between all pairs of networks:

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 267–269, 2016.
DOI: 10.1007/978-3-319-31957-5

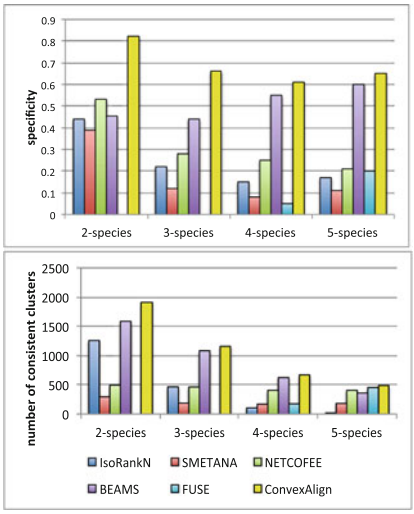


Fig. 1 Specificity and the number of consistent clusters generated by the competing methods for different c on real data where c is the number of species.

$$f_{edge}(\mathcal{A}) = \sum_{1 \le i < j \le N} \vec{1}, y_{ij}, \forall (v_i, v_i') \in E_i, \left(v_j, v_j'\right) \in E_j, 1 \le i < j \le N,$$

where $y_{ij} = X_{ij}\left(v_i, v_j\right) X_{ij}\left(v_i', v_j'\right)$. We aim to find the multiple alignment $\mathcal{A}$ that maximizes a combination of node and edge score as follows: $f = (1-\alpha) f_{node}(\mathcal{A}) + \alpha f_{edge}(\mathcal{A})$, where α describes the trade-off. By doing some calculations, the above objective function can be reformulated as

$$max \sum_{1 \le i < j \le N} (1-\alpha)\left\langle C_{ij}, X_{ij} \right\rangle + \alpha \left\langle \vec{1}, y_{ij} \right\rangle \tag{1}$$

$$y_{ij} \in \{0,1\}^{|E_i| \times |E_j|}, X_{ij} \in \{0,1\}^{|V_i| \times |V_j|}, 1 \le i < j \le N$$

$$B_{ij} y_{ij} \le \mathcal{F}_{ij}\left(X_{ij}\right), \mathrm{X_{ij}} \vec{1} \le \vec{1}, \mathrm{X}_{ij}^{T} \vec{1} \le \vec{1}, 1 \le i < j \le N$$

$$\mathrm{X} \succcurlyeq 0, \mathrm{X_{ii}} = \mathbf{I}_{|\mathrm{V_i}|}, 1 \le i \le N$$

where B_{ij} is coefficient and $\mathcal{F}ij$ is a linear operator that picks the corresponding element of X_{ij} for each constraint.

Optimization via Convex Relaxation. It is NP-hard to directly optimize (1) because the variables are binary. Therefore, we first relax the problem to obtain a convex optimization problem that can be solved to global optimum within polynomial time. We then use an ADMM method to solve the relaxed convex optimization problem that

can align all the proteins together. Finally, a greedy rounding strategy is applied to convert fractional solution to integral.

3 Results

We use both real and synthetic data to evaluate the performance of our method, ConvexAlign, with several popular methods. Tested on the PPI networks of five species human, yeast, fly, mouse and worm, ConvexAlign shows a better performance in terms of specificity and the number of functionally consistent clusters for all the clusters composed of proteins from $c = 2, 3, 4, 5$ species (Fig. 1). We have similar results on synthetic data.

Complexes Detection in Biological Networks via Diversified Dense Subgraphs Mining

Xiuli Ma[1], Guangyu Zhou[2], Jingjing Wang[2], Jian Peng[2], and Jiawei Han[2]

[1] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing, China
xlma@pku.edu.cn

[2] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
{gzhou6,jwang112,jianpeng,hanj}@illinois.edu

1 Introduction

Protein-Protein Interaction (PPI) networks, providing a comprehensive landscape of protein interacting patterns, enable us to explore biological processes and cellular components at multiple resolutions. For a biological process, a number of proteins need to work together to perform the job. Proteins densely interact with each other, forming large molecular machines or cellular building blocks. Identification of such densely interconnected clusters or protein complexes from PPI networks enables us to obtain a better understanding of the hierarchy and organization of biological processes and cellular components.

Most existing methods apply efficient graph clustering algorithms [1–3] on PPI networks, often failing to detect possible densely connected subgraphs and overlapped subgraphs. In this paper, we introduce a novel approximate algorithm to efficiently enumerate putative protein complexes from biological networks. The problem is formulated as finding a diverse set of dense subgraphs that cover as many as proteins as possible. To handle large networks, we take a divide-and-conquer approach to speedup the algorithm in a distributed manner. By comparing with existing clustering-based algorithms on several yeast and human PPI networks, we demonstrate that our method can detect more putative protein complexes and achieve better accuracy.

2 Method

We propose to model the problem of detecting complexes in biological networks as discovering the diversified maximal dense subgraphs. With the density measure, the dense subgraphs, that are protein complexes, can be defined explicitly and flexibly. Instead of enumerating all the dense subgraphs, we only find a small set of diversified maximal dense subgraphs. By maximal, we mean those complexes that are not subset of any other dense subgraphs thus cannot be further extended; By diversified, we mean a diverse set of dense subgraphs which cover

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 270–272, 2016.
DOI: 10.1007/978-3-319-31957-5

as many proteins as possible in the network. Combined into one goal, overlap is allowed but redundancy should be minimized.

In this paper, searching and diversifying are integrated tightly into one whole process. The key component of our algorithm is a set of efficient search trees that compactly traverse all the dense subgraphs by a depth-first construction. A node-specific potential is adopted to guide the search process. Furthermore, we identify two properties, the pseudo anti-monotonicity property for density and the sub-modularity property for diversity, and develop efficient pruning techniques based on these two properties. In this way, we extract the diversified dense subgraphs on the fly during the enumeration of the maximal dense subgraphs, thus greatly improve the scalability of the algorithm. Finally, the algorithm is scaled up in parallel to handle large-scale networks.

3 Result

We extensively evaluate the effectiveness and efficiency of our method on several PPI networks from yeast and human. We first evaluate the number of results and coverage for different methods on all the datasets under different density thresholds, showing our method can detect more complexes, while getting larger coverage. Then, we assess the quality of the predicted complexes by a composite score of three scores: fraction (frac), accuracy (acc) and maximum matching ratio (mmr), on both weighted and unweighted network (which is the binary version of the weighted datasets). In almost all the networks, our approach detects more putative complexes, and achieves higher accuracy and better one-to-one mapping with reference complexes in the ground truth databases than several state-of-art algorithms. The source code and supplementary data are available at https://github.com/zgy921028/MDSMine.

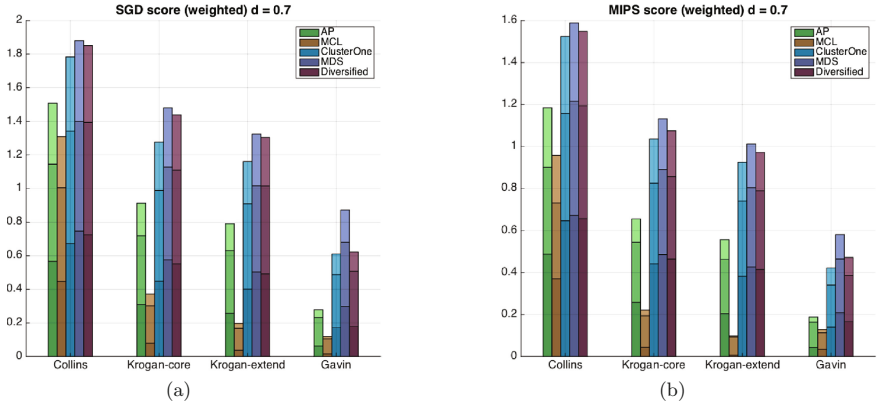


Fig. 1. Results (bottom-up: frac, acc, mmr) of various methods on 4 PPI weighted datasets using SGD (a) and MIPS (b) gold standard.

Acknowledgments. We thank the reviewers for their insightful comments. Xiuli Ma is supported by the National Natural Science Foundation of China under Grant No.61103025 and China Scholarship Council. This work was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1017362, IIS-1320617, and IIS-1354329, HDTRA1-10-1-0120, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC.

References

1. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007)
2. Nepusz, T., Yu, H., Paccanaro, A.: Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods **9**(5), 471–472 (2012)
3. Van Dongen, S.: Graph clustering via a discrete uncoupling process. SIAM J. Matrix Anal. Appl. **30**(1), 121–141 (2008)

Author Index