Short Papers

Targeted Genotyping of Variable Number Tandem Repeats with AdVNTR

Mehrdad Bakhtiari[1(✉)], Sharona Shleizer-Burko[2], Melissa Gymrek[1,2], Vikas Bansal[3], and Vineet Bafna[1]

[1] Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA
mbakhtiari@uscd.edu, vbafna@eng.ucsd.edu
[2] Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[3] Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA

Extended Abstract

Whole Genome Sequencing is increasingly used to identify Mendelian variants in clinical pipelines. These pipelines focus on single nucleotide variants (SNVs) and also structural variants, while ignoring more complex repeat sequence variants. We consider the problem of genotyping *Variable Number Tandem Repeats* (VNTRs), composed of inexact tandem duplications of short (6−100 bp) repeating units. VNTRs span 3% of the human genome, are frequently present in coding regions, and have been implicated in multiple Mendelian disorders (*e.g.*, Medullary cystic kidney disease, Myoclonus epilepsy, and FSHD) and complex disorders such as bipolar disorder. In some cases, the disease associated variants correspond to point mutations in the VNTR sequence while in other cases, changes in the number of tandem repeats (RU count) show a statistical association (or causal relationship) with disease risk. While existing tools are able to recognize VNTR carrying sequence, genotyping VNTRs (determining repeat unit count and sequence variation) from whole genome sequenced reads remains challenging. We describe a method, adVNTR, that models the problems of RU counting and mutation detection using HMMs trained for each target VNTR. adVNTR models can be developed for short-read (Illumina) and single molecule (PacBio) whole genome and exome sequencing. It has three components: (i) HMM training module for model parameter estimation; (ii) read recruitment; and, (iii) estimating RU counts and variant detection. We compared read recruitment with alignment-based methods. The results show that while adVNTR works well for a range of RU counts, other mapping tools work well only when the simulated RU count matches the reference RU count. We performed a long range (LR)PCR experiment on the individual NA12878 to assess the accuracy of the adVNTR genotypes. To test performance of counting of Repeat Units on real data where the true VNTR genotype is not known, we confirmed our results by checking for Mendelian inheritance consistency at 865 VNTRs in two trios.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 243–244, 2018.
https://doi.org/10.1007/978-3-319-89929-9

For short VNTRs, adVNTR can be an effective tool for larger population-scale studies of VNTR genotypes using WGS data replacing labor intensive gel electrophoresis. We found the RU count frequencies for two disease-linked VNTRs in GP1BA and MAOA genes, using 150 PCR-free WGS data. The 2R/3R genotypes in GP1BA are associated with Aspirin Treatment failure for stroke prevention. Notably, our results suggest that the 2R genotype is absent in African populations suggesting that this shorter allele arose after the out of Africa transition. adVNTR is available at https://github.com/mehrdadbakhtiari/adVNTR.

Reference

1. Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V., Bafna, V.: Targeted genotyping of variable number tandem repeats with adVNTR. bioRxiv, p. 221754 (2017)

Positive-Unlabeled Convolutional Neural Networks for Particle Picking in Cryo-electron Micrographs

Tristan Bepler[1,2], Andrew Morin[2,6], Alex J. Noble[3], Julia Brasch[4], Lawrence Shapiro[4,5], and Bonnie Berger[1,2,6(✉)]

[1] Computational and Systems Biology, MIT, Cambridge, MA, USA
bab@mit.edu
[2] Computer Science and AI Laboratory, MIT, Cambridge, MA, USA
[3] National Resource for Automated Molecular Microscopy, Simons Electron Microscopy Center, New York Structural Biology Center, New York, NY, USA
[4] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA
[5] Mortimer B. Zuckerman Mind Brain Behavior Institute, New York, NY, USA
[6] Department of Mathematics, MIT, Cambridge, MA, USA

Background

Structure determination with cryoEM involves reconstructing a 3D molecule from 2D projections. This process often requires tens to hundreds of thousands of experimental projections, or particles. Locating these particles in cryoEM micrographs, referred to as particle picking, is a major bottleneck in the current protein structure determination pipeline. This pipeline generally consists of sample and EM grid preparation, imaging, particle picking, and eventually structure determination. Labeling a sufficient number of particles to determine a high resolution structure can require months of effort – even with the use of existing methods designed to automate the process. Limitations of these tools include high false positive rates, requiring many hand-labeled training examples, and poor performance on non-globular proteins.

In order to better automate particle picking, and thus accelerate structure determination, we newly frame the particle picking problem as an instance of positive-unlabeled classification. In our framework, for a set of micrographs containing particles of interest with a small number labeled for training, we learn a convolutional neural network (CNN) to classify particles from background using a novel generalized-expectation criteria [1] to regularize the model's posterior over the unlabeled micrograph regions. This advance allows us to achieve state-of-the-art particle detection results with minimal hand-labeling required.

B.Berger — This work was partially supported by grants: NIH R01-GM081871, NIH R01-MH1148175, Simons Foundation (349247), NYSTAR, NIH NIGMS (GM103310), the Agouron Institute (F00316) and NIH S10 OD019994-01.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 245–247, 2018.
https://doi.org/10.1007/978-3-319-89929-9

Methods

We develop Topaz, the first particle picking pipeline to use CNNs trained using only positive and unlabeled examples and GE-binomial, a general objective function for learning classifier parameters from positive and unlabeled data. The GE-binomial objective penalizes the negative log-likelihood of the labeled data points while regularizing the classifier's posterior over the unlabeled data to match a binomial distribution prior on the number of unlabeled positives. Denoting the set of labeled positive data points by P, the probabilistic classifier as g, the classifier's posterior over the number of unlabeled positives as q, and the binomial prior as p, the GE-binomial objective function is: $-\mathop{\mathbb{E}}_{x \in P}[\log g(x)] + KL(q \parallel p)$, where KL is the Kullback-Leibler divergence.

In the Topaz pipeline, CNN classifiers are fit to labeled particles and the remaining unlabeled micrograph regions using minibatched stochastic gradient descent to minimize the GE-binomial objective. Predicted particle coordinates are next extracted by scoring each micrograph region with the trained classifier and then using the non-maximum suppression algorithm to greedily select candidate particle coordinates.

Results

We show that the Topaz pipeline is able to accurately detect particles when trained with very few labeled example particles. On the EMPIAR-10096 cryoEM data set [2], Topaz achieves 46% precision at 90% recall with only 1000 labeled particles. In contrast, at the same recall level, EMAN2's byRef method [3] only reaches 33% precision with the same set of labeled particles – corresponding to 71% more false positives than Topaz. Remarkably, Topaz still achieves better precision than EMAN2 at 90% recall with 1/10th and even 1/100th the number of labeled particles. At all numbers of labeled particles tested, we improve substantially over EMAN2's byRef method in area under the precision-recall curve. The relative improvement in particle detection provided by Topaz is even greater on a second, unpublished dataset provided by the Shapiro lab, containing stick-like particles with low signal-to-noise ratio. Furthermore, we show that combining a convolutional decoder with the convolutional feature extractor and classifier learned with GE-binomial to form a hybrid classifier+autoencoder can further improve generalization when very few labeled data points are available. Finally, we demonstrate that our GE-binomial objective function outperforms other positive-unlabeled learning methods never before applied to particle picking. Topaz runs efficiently, training in hours and predicting in seconds with a single consumer grade GPU. We expect Topaz to become an essential component of single particle cryoEM analysis and our GE-binomial objective function to be widely applicable to positive-unlabeled classification problems.

References

1. Mann, G.S., McCallum, A.: Generalized expectation criteria for semi-supervised learning with weakly labeled data. J. Mach. Learn. Res. **11**, 955–984 (2010)
2. Tan, Y.Z., Baldwin, P.R., Davis, J.H., Williamson, J.R., Potter, C.S., Carragher, B., Lyumkis, D.: Addressing preferred specimen orientation in single-particle cryo-EM through tilting. Nat. Methods **14**, 793–796 (2017). https://doi.org/10.1038/nmeth.4347
3. Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., Ludtke, S.J.: EMAN2: an extensible image processing suite for electron microscopy. J. Struct. Biol. **166**, 205–213 (2007). https://doi.org/10.1016/j.jsb.2006.05.009

Designing RNA Secondary Structures Is Hard

Édouard Bonnet[1], Paweł Rzążewski[2], and Florian Sikora[3(✉)]

[1] Department of Computer Science, Middlesex University, London, UK
edouard.bonnet@dauphine.fr
[2] Faculty of Mathematics and Information Science,
Warsaw University of Technology, Warsaw, Poland
p.rzazewski@mini.pw.edu.pl
[3] Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE,
Paris, France
florian.sikora@dauphine.fr

An RNA sequence is a word over an alphabet on four elements $\{A, C, G, U\}$ called bases. RNA sequences fold into secondary structures where some bases pair with one another while others remain unpaired. Pseudoknot-free secondary structures can be represented as well-parenthesized expressions with additional dots, where pairs of matching parentheses symbolize paired bases and dots, unpaired bases. The two fundamental problems in RNA algorithmic are to *predict* how sequences fold within some model of energy and to *design* sequences of bases which will fold into targeted secondary structures. Predicting how a given RNA sequence folds into a pseudoknot-free secondary structure is known to be solvable in cubic time since the eighties [15, 16] and in truly subcubic time by a recent result of Bringmann et al. [3], whereas Lyngsø has shown it is NP-complete if pseudoknots are allowed [13]. As a stark contrast, it is unknown whether or not designing a given RNA secondary structure is a tractable task; this has been raised as a challenging open question by several authors [2, 6, 7, 9, 11, 14]. Because of its crucial importance in a number of fields such as pharmaceutical research and biochemistry, there are dozens of heuristics and software libraries dedicated to RNA secondary structure design [1, 2, 4, 5, 8]. It is therefore rather surprising that the computational complexity of this central problem in bioinformatics has been unsettled for decades.

As our main result we show that, in the simplest model of energy which is the Watson-Crick model the design of secondary structures is NP-complete if one adds natural constraints of the form: *index i of the sequence has to be labeled by base b*. This negative result suggests that the same lower bound holds for more realistic models of energy. It is noteworthy that the additional constraints are by no means artificial: they are provided by all the RNA design pieces of software and they do correspond to the actual practice (see for example the instances of the EteRNA project [12]). Our reduction from a variant of 3-Sat has as main ingredients: arches of parentheses of different widths, a linear order interleaving variables and clauses, and an intended *rematching strategy* which increases the number of pairs if and only if the three literals of a same clause are false. The correctness of the construction is also quite intricate; it relies on the polynomial

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 248–250, 2018.
https://doi.org/10.1007/978-3-319-89929-9

algorithm for the design of saturated structures – secondary structures without dots – by Haleš et al. [9, 10], counting arguments, and a concise case analysis.

We also show that a naive brute-force algorithm for RNA Design can be improved by a careful structural analysis.

References

1. Aguirre-Hernández, R., Hoos, H.H., Condon, A.: Computational RNA secondary structure design: empirical complexity and improved methods. BMC Bioinf. **8**(1), 34 (2007)
2. Andronescu, M., Fejes, A.P., Hutter, F., Hoos, H.H., Condon, A.: A new algorithm for RNA secondary structure design. J. Mol. Biol. **336**(3), 607–624 (2004)
3. Bringmann, K., Grandoni, F., Saha, B., Williams, V.V.: Truly sub-cubic algorithms for language edit distance and RNA-folding via fast bounded-difference min-plus product. In: IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, pp. 375–384 (2016)
4. Butterfoss, G.L., Kuhlman, B.: Computer-based design of novel protein structures. Annu. Rev. Biophys. Biomol. Struct. **35**, 49–65 (2006)
5. Churkin, A., Retwitzer, M. D. Reinharz, V., Ponty, Y., Waldispühl, J., Barash. D.: Design of RNAs: comparing programs for inverse RNA folding. Briefings Bioinf. (2017)
6. Condon, A.: Problems on RNA secondary structure prediction and design. In: Baeten, J.C.M., Lenstra, J.K., Parrow, J., Woeginger, G.J. (eds.) ICALP 2003. LNCS, vol. 2719, pp. 22–32. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45061-0_2
7. Condon, A.: RNA molecules: glimpses through an algorithmic lens. In: Correa, J.R., Hevia, A., Kiwi, M. (eds.) LATIN 2006. LNCS, vol. 3887, pp. 8–10. Springer, Heidelberg (2006). https://doi.org/10.1007/11682462_2
8. García-Martín, J. A., Clote, P., Dotú, I.: RNAiFold: a web server for RNA inverse folding and molecular design. Nucleic Acids Res. **41**(Webserver-Issue), 465–470 (2013)
9. Hales, J., Héliou, A., Manuch, J., Ponty, Y., Stacho, L.: Combinatorial RNA design: designability and structure-approximating algorithm in watson-crick and nussinov-jacobson energy models. Algorithmica **79**(3), 835–856 (2017)
10. Haleš, J., Maňuch, J., Ponty, Y., Stacho, L.: Combinatorial RNA design: designability and structure-approximating algorithm. In: Cicalese, F., Porat, E., Vaccaro, U. (eds.) CPM 2015. LNCS, vol. 9133, pp. 231–246. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19929-0_20
11. Jedwab, J., Petrie, T., Simon, S.: An infinite class of unsaturated rooted trees corresponding to designable RNA secondary structures. CoRR, abs/1709.08088 (2017)
12. Lee, J., Kladwang, W., Lee, M., Cantu, D., Azizyan, M., Kim, H., Limpaecher, A., Gaikwad, S., Yoon, S., Treuille, A., Das, R., Participants, E.R.N.A.: RNA design rules from a massive open laboratory. Proc. Nat. Acad. Sci. **111**(6), 2122–2127 (2014)
13. Lyngsø, R.B.: Complexity of pseudoknot prediction in simple models. In: Díaz, J., Karhumäki, J., Lepistö, A., Sannella, D. (eds.) ICALP 2004. LNCS, vol. 3142, pp. 919–931. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-27836-8_77

14. Lyngsø, R.B.: Inverse folding of RNA (2012). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.226.5439&rep=rep1&type=pdf
15. Nussinov, R., Jacobson, A.B.: Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc. Nat. Acad. Sci. **77**(11), 6309–6313 (1980)
16. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. **9**(1), 133–148 (1981)

Generalizable Visualization of Mega-Scale Single-Cell Data

Hyunghoon Cho[1], Bonnie Berger[1,2(✉)], and Jian Peng[3(✉)]

[1] CSAIL, MIT, Cambridge, MA 02139, USA
[2] Department of Mathematics, MIT, Cambridge, MA 02139, USA
bab@mit.edu
[3] Department of Computer Science, UIUC, Urbana, IL 61801, USA
jianpeng@illinois.edu

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has been a key tool in dissecting intercellular variation in biomedical sciences. A standard analysis for scRNA-seq data is to visualize the cells in a low-dimensional (2D or 3D) space via methods such as *t-stochastic neighbor embedding* (t-SNE) [1], where each cell is represented as a dot and dots of cells with similar expression profiles are located close to each other in space. Such visualization reveals the salient structure of the data in a form that is easy for researchers to grasp and further analyze.

Recent advances in sequencing technologies has led to an exponential growth in the number of cells sequenced in a study. For example, 10x Genomics recently published a dataset of 1.3 million mouse neurons [2]. The emergence of such *mega*-scale data poses new computational challenges before they can be widely adopted, as many of the existing tools for scRNA-seq analysis (including t-SNE) require prohibitive runtimes or computational resources for data of this size.

We introduce *neural t-SNE* (net-SNE), a scalable and generalizable method for visualizing millions of cells for scRNA-seq analysis. net-SNE learns a high-quality mapping *function* that takes an expression profile as input and outputs a low-dimensional embedding in 2D or 3D for visualization. Unlike t-SNE, the mapping function learned by net-SNE can be used to map *previously unseen* cells. In addition to allowing fast visualization of datasets with millions of cells, net-SNE enables novel workflows for single-cell genomics, where newly observed cells are visualized in the context of existing datasets for translational analysis.

2 Methods

Our method (net-SNE) models the position of each cell in the visualization as the output of a parameterized map evaluated at the given expression profile. We use feedforward neural networks (NNs) to represent the embedding function, drawing from the intuition that NNs have sufficient expressive capacity to find high-quality maps similar to those typically uncovered by t-SNE. To optimize

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 251–253, 2018.
https://doi.org/10.1007/978-3-319-89929-9

the NN parameters, net-SNE minimizes the same objective score optimized by t-SNE via gradient descent. This choice of objective allows net-SNE to emulate the behavior of t-SNE while newly achieving generalizability and scalability. Notably, net-SNE is compatible with existing optimizations for t-SNE—our implementation of net-SNE incorporates an efficient variant of t-SNE based on Barnes-Hut approximation [1]. We achieve further efficiency by employing stochastic optimization techniques, where only a subset of cells are used to approximate each parameter update. Such stochastic acceleration is newly enabled by net-SNE due to the fact that parameters being optimized are *shared* across all cells.

3 Results

We observed that net-SNE learns an embedding that closely matches t-SNE on 13 scRNA-seq datasets with known clusters in terms of both visual quality and clustering accuracy. Furthermore, when an entire cluster of cells was withheld and placed onto the visualization after the fact, net-SNE accurately positioned the held-out cells as a distinct cluster, despite not having seen any cells from the missing cluster. To demonstrate fast visualization of mega-scale datasets, we also pre-trained net-SNE on a random subset of 100K cells from the 10x Genomics dataset and used the learned embedding to instantly visualize the entire dataset in less than a minute. This approach obtained a higher quality map than t-SNE with the default parameters, the latter of which took 13 h to finish. While the pre-training of net-SNE took 3 h in our experiment, we note that a pre-trained embedding may be readily available in certain use cases. We provide example visualizations by net-SNE in Fig. 1.

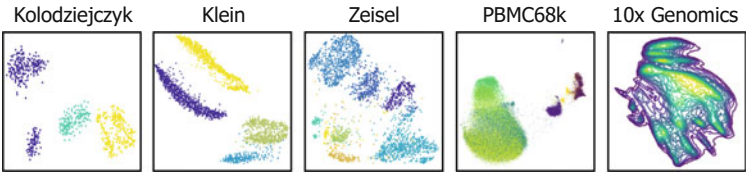

Fig. 1. Example 2D visualizations of single-cell RNA-seq datasets by net-SNE

Overall, our results demonstrate that net-SNE not only learns high quality maps like t-SNE, but also gracefully generalizes to unseen cells. This allows net-SNE to efficiently visualize mega-scale single-cell data by using a pre-trained embedding from a subsampled or an existing dataset. Our work is widely applicable to other data science domains with millions of data points to be visualized.

Acknowledgements. This work was partially supported by NIH R01GM081871.

References

1. Van Der Maaten, L.: Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. **15**(1), 3221–3245 (2014)
2. 10x Genomics: Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium™ Single Cell 3' Solution. Application Note (2017). https://www.10xgenomics.com/single-cell/. Accessed October 2017

Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis

G. Durif[1,2(✉)], L. Modolo[1,3,4], J. E. Mold[4], S. Lambert-Lacroix[5], and F. Picard[1]

[1] LBBE, UMR CNRS 5558, Université Lyon 1, 69622 Villeurbanne, France
ghislain.durif@inria.fr
[2] Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[3] LBMC UMR 5239 CNRS/ENS Lyon, 69007 Lyon, France
[4] Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden
[5] UMR 5525 Université Grenoble Alpes/CNRS/TIMC-IMAG, 38041 Grenoble, France

The combination of massive parallel sequencing with high-throughput cell biology technologies has given rise to single-cell Genomics. Similar to the paradigm shift of the 90 s characterized by the first molecular profiles of tissues, it is now possible to characterize molecular heterogeneities at the cellular level (Saliba et al. 2014). The statistical characterization of heterogeneities in single-cell expression data thus requires an appropriate model, since the transcripts abundance is quantified for each cell using read counts. Hence, standard methods based on Gaussian assumptions are likely to fail to catch the biological variability of lowly expressed genes, and Poisson or Negative Binomial distributions constitute an appropriate framework (Chen et al. 2016). Moreover, dropouts, either technical (due to sampling difficulties) or biological (no expression or stochastic transcriptional activity), constitute another major source of variability in scRNA-seq (single-cell RNA-seq) data, which has motivated the development of the so-called Zero-Inflated models (Kharchenko et al. 2014). A standard and popular way of quantifying and visualizing the variability within a dataset is dimension reduction, principal component analysis (PCA) being the most widely used technique in practice. Model-based PCA (Collins et al. 2001) offers the unique advantage to be adapted to the data distribution and to be based on an appropriate metric, the Bregman divergence. It consists in specifying the distribution of the data through a statistical model. A probabilistic zero-inflated version of the Gaussian PCA was proposed by Pierson and Yau (2015) in the context of single cell data analysis (the ZIFA method). However, scRNA-seq data may be better analyzed by methods dedicated to count data such as the Non-negative Matrix Factorization (Lee and Seung 1999, NMF) or the Gamma-Poisson factor model (Cemgil 2009). However, none of the currently available dimension reduction methods fully model single-cell expression data, characterized by overdispered zero inflated counts (Zappia et al. 2017). Our method is based on a probabilistic count matrix factorization (pCMF). We propose a dimension reduction method that is dedicated to over-dispersed counts

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 254–255, 2018.
https://doi.org/10.1007/978-3-319-89929-9

with dropouts, in high dimension. Our factor model takes advantage of the Poisson Gamma representation to model counts from scRNA-seq data (Zappia et al. 2017). In particular, we use Gamma priors on the distribution of principal components. We model dropouts with a Zero-Inflated Poisson distribution, and we introduce sparsity in the model thanks to a spike-and-slab approach (Malsiner-Walli and Wagner 2011) that is based on a two component sparsity-inducing prior on loadings (Titsias and Lázaro-Gredilla 2011). The model is inferred using a variational EM algorithm that scales favorably to data dimension, as compared with Markov Chain Monte Carlo (MCMC) methods (Blei et al. 2017). Then we propose a new criterion to assess the quality of fit of the model to the data, as a percentage of explained deviance, because the standard variance reduction that is used in PCA needs to be adapted to the new framework dedicated to counts. We show that pCMF better catches the variability of simulated data and experimental scRNA-seq datasets. Finally, pCMF is available in the form of a R package available at https://gitlab.inria.fr/gdurif/pCMF.

References

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. (2017). (just-accepted)

Cemgil, A.T.: Bayesian inference for nonnegative matrix factorisation models. Computational Intelligence and Neuroscience (2009)

Chen, H.-I.H., Jin, Y., Huang, Y., Chen, Y.: Detection of high variability in gene expression from single-cell RNA-seq profiling. BMC Genomics **17**(Suppl 7) (2016)

Collins, M., Dasgupta, S., Schapire, R.E.: A generalization of principal components analysis to the exponential family. In: Advances in Neural Information Processing Systems, pp. 617–624 (2001)

Kharchenko, P.V., Silberstein, L., Scadden, D.T.: Bayesian approach to single-cell differential expression analysis. Nat. Methods **11**(7), 740 (2014)

Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)

Malsiner-Walli, G., Wagner, H.: Comparing spike and slab priors for Bayesian variable selection. Austrian J. Stat. **40**(4), 241–264 (2011)

Pierson, E., Yau, C.: ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. **16**, 241 (2015)

Saliba, A.-E., Westermann, A.J., Gorski, S.A., Vogel, J.: Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. **42**(14), 8845–8860 (2014)

Titsias, M. K., & Lázaro-Gredilla, M.: Spike and slab variational inference for multitask and multiple kernel learning. In: Advances in Neural Information Processing Systems, pp. 2339–2347 (2011)

Zappia, L., Phipson, B., Oshlack, A.: Splatter: simulation of single-cell RNA sequencing data. Genome Biol. **18**, 174 (2017)

Fixed-Parameter Tractable Sampling for RNA Design with Multiple Target Structures

Stefan Hammer[1,2,3], Yann Ponty[4,5(✉)], Wei Wang[4,5], and Sebastian Will[2]

[1] Department of Computer Science and Interdisciplinary Center for Bioinformatics, University Leipzig, 04107 Leipzig, Germany
[2] Department of Theoretical Chemistry, Faculty of Chemistry, University of Vienna, 1090 Vienna, Austria
[3] Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, 1090 Vienna, Austria
[4] CNRS UMR 7161 LIX, Ecole Polytechnique, Bat. Turing, 91120 Palaiseau, France
[5] AMIBio team, Inria Saclay, Bat Alan Turing, 91120 Palaiseau, France
yann.ponty@lix.polytechnique.fr

Motivation. Engineering artificial biological systems promises broad applications in synthetic biology, biotechnology and medicine. Here, the rational design of multi-stable RNA molecules is especially powerful, since RNA can be generated with highly specific properties and programmable functions. In particular, designing artificial riboswitches became popular due to their potential as versatile biosensors [1]. Effective in-silico methods proved to greatly facilitate the design approach and have tremendous impact on their cost and feasibility.

Statement of Problem. Most methods for computational design share a similar overall strategy: one or several initial seed sequences are generated and optimized subsequently. In this contribution we revisit the first main ingredient of (multi-target) design methods, namely the sampling of sequences, which energetically favor several given target structures at the same time. While previous multi-target methods [4, 6] relied on *ad-hoc* sampling strategies, sampling seeds from the uniform distribution was solved only recently [2, 3].

Algorithmic Contributions. We generalize Boltzmann sampling for RNA design, which was recently shown powerful for single targets in `IncaRNAtion` [5], to design for multiple structural targets. After showing that even uniform sampling is #P-hard, we introduce the tree decomposition-based fixed parameter tractable (FPT) sampling algorithm `RNARedPrint`. Finally, we combine our FPT stochastic sampling algorithm with multi-dimensional Boltzmann sampling over distributions controlled by expressive RNA energy models. We show that sampling t sequences of length n for k target structures takes $\mathcal{O}(2^d\, n\, k + t\, n\, k)$ time, where $d := \min(w + c + 1, 2(w + 1))$, depending on the tree width w of the dependency graph (covering all dependencies between sequence positions introduced by the energy function) as well as the number c of connected components in the compatibility graph (covering the constraints enforcing canonical base pairings). Due

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 256–258, 2018.
https://doi.org/10.1007/978-3-319-89929-9

to a constraint framework, `RNARedPrint` supports generic Boltzmann-weighted sampling for arbitrary additive RNA energy models; this moreover enables targeting specific free energies or GC-content, compare Fig. 1.

Empirical Results. We study general properties of the approach and generate biologically relevant multi-target Boltzmann-weighted designs. Thereby, we observe significant improvements over ad-hoc methods or even uniform sampling.

Extensibility of the Approach. The presented framework is designed to enable even more general new possibilities for sequence generation in the field of RNA sequence design by enforcing additional constraints, including more complex sequence constraints, e.g. forbidden motifs in the designed sequences.

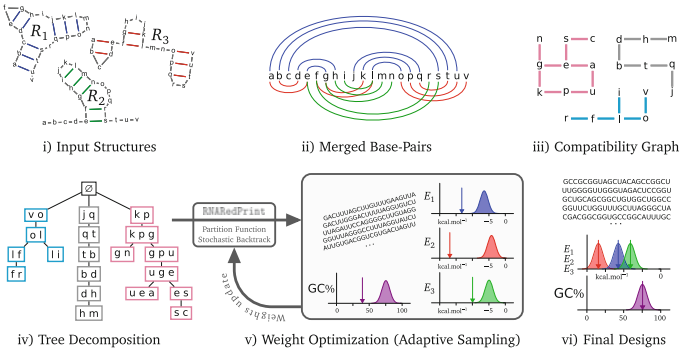

Fig. 1. General outline of `RNARedPrint`. From a set of target secondary structures (i), base-pairs are merged (ii) into a compatibility graph (iii). Based on its tree decomposition (iv), we compute the partition function, followed by a Boltzmann sampling of valid sequences (v). An adaptive scheme learns weights to achieve targeted energies and GC-content, leading to the production of suitable designs (vi).

Availability as free software: https://github.com/yannponty/RNARedPrint

References

1. Domin, G., Findeiß, S., Wachsmuth, M., Will, S., Stadler, P.F., Mörl, M.: Applicability of a computational design approach for synthetic riboswitches. Nucleic Acids Res. **45**(7), 4108–4119 (2017)
2. Hammer, S., Tschiatschek, B., Flamm, F., Hofacker, I.L., Findeiß, S.: RNAblueprint: flexible multiple target nucleic acid sequence design. Bioinformatics **33**, 2850–2858 (2017)
3. Höner zu Siederdissen, C., Hammer, S., Abfalter, I., Hofacker, I.L., Flamm, C., Stadler, P.F.: Computational design of RNAs with complex energy landscapes. Biopolymers **99**, 1124–1136 (2013)

4. Lyngso, R.B., Anderson, J.W.J., Sizikova, E., Badugu, A., Hyland, T., Hein, J.: Frnakenstein: multiple target inverse RNA folding. BMC Bioinf. **13**, 260 (2012)
5. Reinharz, V., Ponty, Y., Waldispühl, J.: A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution. Bioinformatics **29**, i308–i315 (2013)
6. Taneda, A.: Multi-objective optimization for RNA design with multiple target secondary structures. BMC Bioinf. **16**, 280 (2015)

Contribution of Structural Variation to Genome Structure: TAD Fusion Discovery and Ranking

Linh Huynh[1] and Fereydoun Hormozdiari[1,2,3](✉)

[1] Genome Center, UC Davis, Davis, USA
[2] MIND Institute, UC Davis, Davis, USA
[3] Biochemistry and Molecular Medicine, UC Davis, Davis, USA
fhormozd@ucdavis.edu

Introduction

The significant contribution of structural variants (e.g. deletion, insertion, and inversion) to function, disease, and evolution is well reported. However, in many cases, the mechanism by which these variants contribute to the phenotype is not well understood. This is especially the case for studying non-coding structural variants and their potential biological impact. With the advent of high-throughput chromosome conformation capture (Hi-C [1]) we have novel insights into genome structure and its contribution to gene regulation. Using Hi-C data we are able to study the genomic interactions, such as enhancer-promoter interactions that are the main mechanism for gene regulation. The analysis of Hi-C data has also provided evidence that genome folds into different compartments and domains which guide the regions of the genome that can interact with each other. One of these types of domains discovered is called topological associated domains (TADs) and has provided a novel understanding of how genome structure contributes to regulation [2]. Recent studies reported structural variants (SVs) that disrupted the three-dimensional genome structure by fusing two TADs, such that enhancers from one TAD interacted with genes from the other TAD, could cause severe developmental disorders [3]. However, no method exists for directly scoring and ranking structural variations based on their effect on the three-dimensional structure such as the TAD disruption. In this paper, we formally define TAD fusion and provide a combinatorial approach for assigning a score to quantify the level of TAD fusion for each deletion denoted as TAD fusion score.

Methods

Our goal is to develop a computational method that can provide a score for deletions based on its level of modifying the 3D genomic structure and potential of causing a TAD fusion. In our method, the input consists of a Hi-C contact matrix of the genome with reference allele (i.e., without the deletion) and the

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 259–260, 2018.
https://doi.org/10.1007/978-3-319-89929-9

coordinates of the deletion. The output is a score representing the number of new genomic interactions made (i.e., TAD fusion score) as a result of the deletion. For this paper, we are only considering deletions, however, this approach can be extended to consider other SV types (e.g. translocations).

We propose a two-step framework for calculating the TAD fusion score: (i) predicting a new Hi-C contact matrix G of the mutated chromosome (i.e. with the deletion) given the Hi-C contact matrix H of a genome without the deletion and the deletion coordinates as the inputs; (ii) comparing this predicted/new Hi-C contact matrix G with the original Hi-C contact matrix H to estimate the number of new interactions created as a result of that deletion. For the first step, we extend the power law model (i.e. length-based model) by adding new parameters that represent the TAD structure. By that, all model parameter values can be estimated by solving a linear programming. For the second step, we define TAD fusion score as the expected number of additional genomic interactions created as a result of the deletion. Here, the genomic interactions can be defined by a simple step function or by a Bayesian formula.

Results

We show that our extended model gives a better prediction of the Hi-C contact matrix than the (length-based) power law model. In addition, our method can accurately score deletions which result in TAD fusion, and it outperforms the approaches which use predicted TADs to overlay the deletion on them for predicting TAD fusion. Furthermore, we show that our method correctly gives higher scores to deletions reported to cause developmental disorders as a result of disrupting genome structure in comparison to the deletions reported in the 1000 genomes project. Finally, we also show that deletions that cause TAD fusion are rare and under negative selection in general population.

TAD fusion score is available at https://github.com/huynhvietlinh/FusionScore.

References

1. Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326**(5950), 289–293 (2009)
2. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**(7398), 376–380 (2012)
3. Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al.: Disruptions of topological chromatin domains cause pathogenic rewiring ofgene-enhancer interactions. Cell **161**(5), 1012–1025 (2015)

Assembly of Long Error-Prone Reads Using Repeat Graphs

Mikhail Kolmogorov[1(✉)], Jeffrey Yuan[2], Yu Lin[3], and Pavel Pevzner[1]

[1] Department of Computer Science and Engineering, University of California, San Diego, La Jolla, USA
mkolmogo@ucsd.edu

[2] Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, USA

[3] Research School of Computer Science, Australian National University, Canberra, Australia

The problem of genome assembly is ultimately linked to the r*epeat characterization problem*, the compact representation of all repeat families in a genome as a *repeat graph* [1]. Long read technologies have not made the repeat characterization problem irrelevant. Instead, they have simply shifted the focus from short repeats to longer repeats comparable in length to the median SMS read size; e.g., Kamath et al. [2] analyzed many bacterial genomes that existing SMS assemblers failed to assemble into a single contig. Since even bacterial (let alone, eukaryotic) genomes have long repeats, SMS assemblers currently face the same challenge that short read assemblers faced a decade ago, albeit at a different scale of repeat lengths.

Most algorithms for assembling long error-prone reads use an *overlap-layout-consensus (OLC)* approach that does not provide a repeat characterization [3, 4]. In contrast, *de Bruijn graphs* emerged as a popular approach for short read assembly because they offered an elegant representation of all repeats in a genome that reveals their mosaic structure. Most short read assemblers construct the de Bruijn graph based on all k-mers in reads and further transform it into an *assembly graph* using various *graph simplification* procedures. However, in the case of SMS reads, the key assumption of the de Bruijn graph approach (that most k-mers from the genome are preserved in multiple reads) does not hold even for short k-mers, let alone for long k-mers (e.g., $k = 1000$). As a result, various issues that have been addressed in short read assembly (e.g., how to deal with the fragmented de Bruijn graph, how to transform it into an assembly graph, etc.) remain largely unaddressed in the case of the de Bruijn graph approach to SMS assemblies.

Here, we describe the Flye algorithm for constructing repeat graphs (which have properties similar to de Bruijn graphs) from SMS reads. Flye is built on top of the ABruijn assembler [5], which generates *accurate* overlapping contigs but does not reveal the repeat structure of the genome. In contrast to ABruijn, Flye initially generates inaccurate overlapping contigs (i.e., contigs with potential assembly errors representing random walks on the true repeat graph) and combines these *initial* contigs into an accurate assembly graph that encodes all possible assemblies consistent with the reads. Flye further resolves *bridged* repeats

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 261–262, 2018.
https://doi.org/10.1007/978-3-319-89929-9

in the assembly graph thus constructing a new, less tangled assembly graph, and finally outputs accurate *final* contigs formed by paths in this graph. Flye also introduces a new algorithm that uses small differences between repeat copies to resolve *unbridged* repeats that are not spanned by any reads. We benchmarked Flye against several state-of-the-art SMS assemblers using various datasets and demonstrated that it generates accurate assemblies while also providing insight into how to plan additional experiments (e.g., using contact or optical maps) to finish the assembly. Flye is freely available at http://github.com/fenderglass/Flye.

References

1. Pevzner, P.A., Tang, H., Tesler, G.: De novo repeat classification and fragment assembly. Genome Res. **14**(9), 1786–1796 (2004)
2. Kamath, G.M., Shomorony, I., Xia, F., Courtade, T.A., Tse, N.: D: HINGE: long-read assembly achieves optimal repeat resolution. Genome Res. **27**(5), 747–756 (2017)
3. Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G.R.: Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods **13**(12), 1050 (2016)
4. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M.: Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. **27**(5), 722–736 (2017)
5. Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M., Pevzner, P.A.: Assembly of long error-prone reads using de Bruijn graphs. Proc. Nat. Acad. Sci. **113**(52), E8396–E8405 (2016)

A Multi-species Functional Embedding Integrating Sequence and Network Structure

Mark D. M. Leiserson[1], Jason Fan[1], Anthony Cannistra[2], Inbar Fried[3], Tim Lim[4], Thomas Schaffner[5], Mark Crovella[4], and Benjamin Hescott[6(✉)]

[1] Department of Computer Science, University of Maryland, College Park, USA
[2] Department of Biology, University of Washington, Seattle, USA
[3] University of North Carolina Medical School, Chapel Hill, USA
[4] Department of Computer Science, Boston University, Boston, USA
[5] Department of Computer Science, Princeton University, Princeton, USA
[6] College of Computer and Information Science, Northeastern University, Boston, USA
b.hescott@northeastern.edu

Introduction. Transferring biological knowledge between species is fundamental for many important problems in genetics. These problems range from the molecular-level, such as predicting protein function or genetic interactions [4], to the organism-level, such as predicting human disease models [5]. The most common approach researchers have taken is to use orthologs inferred from DNA sequencing data. More recently, researchers have sought to expand beyond sequence-based orthologs using high-throughput proteomics data under the hypothesis that genes with similar topology in protein-protein interaction (PPI) networks have similar functions. Many methods have been introduced to infer homology across species (i.e. a node matching) from sequence similarity and PPI networks, including network alignment [1]. More recently, Jacunski, et al. [4] identified *connectivity homologous* gene pairs using a small set of features derived from PPI networks. These prior works are focused on node matching and constructing node feature vectors, but do not address the problem of embedding genes from different species into a shared, general-purpose space.

Methods. We introduce a new algorithm, Homology Assessment across Networks using Diffusion and Landmarks (HANDL), that leverages graph kernels to embed nodes from two PPI networks into a biologically meaningful and general-purpose vector space using network and sequence data.[1] Kernels, particularly kernels that capture random walks and/or heat diffusion processes on graphs, have been widely and successfully used for computing similarity between nodes within biological networks [2].

The main computational challenge HANDL solves is relating network kernel matrices from different species. Because the kernel matrices from networks of

[1] An implementation of HANDL is available at https://github.com/lrgr/HANDL.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 263–265, 2018.
https://doi.org/10.1007/978-3-319-89929-9

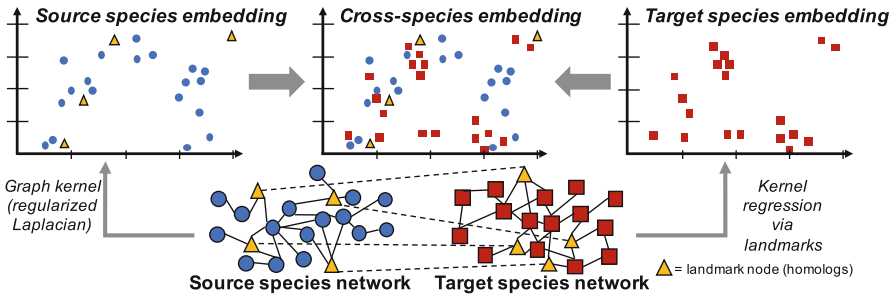

Fig. 1. HANDL embeds nodes into a shared vector space.

different species have different dimensions, traditional kernel transfer learning approaches (e.g. [3]) cannot be directly applied. We show a schematic of the HANDL algorithm in Fig. 1. HANDL takes as input a *source* network, a *target* network, and a set of *landmarks* shared between the networks to embed nodes from the target species into the vector space of the source species. The inner-product between embeddings gives *HANDL similarity scores* between nodes in different species. As HANDL is a general algorithm, the landmarks and graph kernel can be customized for particular applications. In this work, we use a subset of homologs between the source and target species as landmarks and the regularized Laplacian kernel specifically to capture protein functional similarity.

Results. We show that the human-mouse and baker's-fission yeast cross-species embeddings constructed by HANDL are biologically meaningful with three cross-species tasks. First, we find that HANDL similarity scores are strongly correlated with cross-species functional similarity, and that pairs with the highest HANDL similarity scores are more functionally similar than pairs with the closest connectivity homology profiles [4]. Next, we use the algorithm and data from McGary, et al. [5] and *HANDL-homologs* (node pairs with high HANDL similarity scores) to find new, novel human-mouse disease models (phenologs, i.e orthologous phenotypes) that are supported by biological literature. Finally, we show that node vectors themselves are of more general use. We use HANDL to transfer knowledge of synthetic lethal (SL) interactions in baker's to fission yeast (and vice versa). We compute embeddings for the source and target species then train a support vector machine (SVM) only on embeddings of the source species. We find that that the SVM also separates embeddings of the target species with respect to SLs and non-SLs on previously unseen data.

These results show how HANDL can transfer knowledge of genetics between humans and model organisms. We anticipate that HANDL can serve as the foundation for more sophisticated approaches for transfer learning across species.

References

1. Clark, C., Kalita, J.: A comparison of algorithms for the pairwise alignment of biological networks. Bioinformatics **30**(16), 2351–2359 (2014)
2. Cowen, L., Ideker, T., Raphael, B.J., Sharan, R.: Network propagation: a universal amplifier of genetic associations. Nat. Rev. Genet. (2017)
3. Huang, J., Smola, A.J., Gretton, A., Borgwardt, K. M., and Scholkopf, B.: Correcting sample selection bias by unlabeled data. Adv. Neural Inf. Process. Syst. 601–608 (2006)
4. Jacunski, A., Dixon, S.J., Tatonetti, N.P.: Connectivity homology enables interspecies network models of synthetic lethality. PLoS Comp. Bio. **11**(10), e1004506 (2015)
5. McGary, K.L., Park, T., Woods, J.O., Cha, H., Wallingford, J.B., Marcotte, E.M.: Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc. Natl. Acad. Sci. 107(14), 6544–6549 (2010)

Deciphering Signaling Specificity with Deep Neural Networks

Yunan Luo[1], Jianzhu Ma[2], Yang Liu[1], Qing Ye[1], Trey Ideker[2], and Jian Peng[1(✉)]

[1] Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, USA
[2] School of Medicine, University of California San Diego, La Jolla, USA
jianpeng@illinois.edu

1 Introduction

Protein kinase phosphorylation is one of the primary forms of post-translation modification (PTM) that transduce cellular signals and regulate cellular processes. Defective signal transductions, which are associated with protein phosphorylation, have been linked to many human diseases, such as cancer. Defining the organization of the phosphorylation-based signaling network and, in particular, identifying kinase-specific substrates can help reveal the molecular mechanism of the signaling network and understand their impacts on human diseases.

2 Methods

We present DeepSignal, a deep learning based method for predicting the substrate specificity of kinase domains. Unlike most of the previous methods that only focus on using substrate sequences to derive the kinases specificity, DeepSignal takes into account the information in both kinase domain sequences and substrate peptides, and translates a kinase sequences into its specificity profile (e.g., a position-specific scoring matrix, PSSM). DeepSignal employs the Long Short-Term Memory (LSTM) network, a deep learning architecture with memory units, to process the kinase sequences with various lengths using a single model, enabling the learning of universal knowledge across multiple kinase domains. Our deep learning based method is able to automatically extract complex features in kinase domain sequences that best explains the substrate specificity of this kinase. For example, with the memory ability of LSTM, DeepSignal can exploit and record the long and short range dependencies between residues spanning over an arbitrary distance in the kinase domain, which is challenging for previous non-deep learning methods of phosphosites prediction. In addition, DeepSignal can transfer the knowledge from currently available kinase-substrate data to predict phosphosites for new kinases, which is infeasible for many existing kinase-specific methods.

Y. Luo, J. Ma, and Y. Liu — Equal contribution.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 266–268, 2018.
https://doi.org/10.1007/978-3-319-89929-9

3 Results

We evaluated the ability of DeepSignal on predicting the substrate specificity of kinase domains. Our method is able to achieve 0.875 AUROC (area under the receiver operating characteristic curve) and 0.21 AUPRC (area under the precision-recall curve) scores in a five-fold cross-validation, which is a substantial improvement over previous methods GPS 2.0 [1] and NetPhorest [2]. To test the generalization ability of our method, we further apply DeepSignal to predict the binding specificity of SH2 domain (Fig. 1), another phosphorylation-based signaling modular domain, on four high-throughput datasets. DeepSignal significantly outperforms two SH2-peptide interaction methods (SMALI [3] and SH2PepInt [4]) and one general protein-protein interaction method (PrePPI [5]). Although trained on 80% of the data in the five-fold cross-validation, our method still achieves higher or comparable AUROC scores when compared to a method (MSM/D-PEM [6]) that was pre-trained on all the binding data of each dataset. Overall, these results demonstrated the ability of DeepSignal on predicting the binding specificity of phosphorylation-based signaling domains.

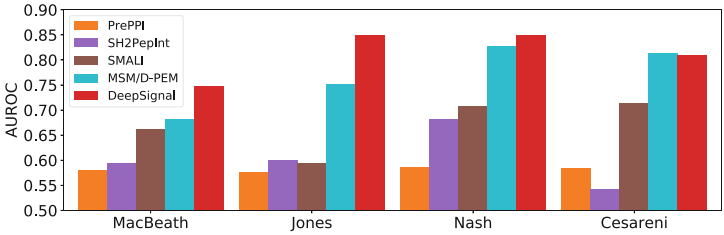

Fig. 1. Evaluation of prediciton performance on prediction of the binding between SH2 domains and phosphotyrosine peptides.

To study the impact of mutations on cancer, we used DeepSignal to construct the signaling network using only the protein primary sequences of 16,254 proteins, including 307 kinase domains, 122 SH2 domains and 190,427 phosphoproteins across 18 cancer types. For each cancer type, we mapped all the coding mutations from TCGA on the protein sequences. This resulted 6,286 mutations on kinase domains, 776 mutations on SH2 domains and 37,996 mutations on phosphoproteins. We use DeepSignal to quantify the change of the binding specificity caused by the cancer mutations of a given kinase/SH2-peptide, and predict a ranking list of single-nucleotide variants (SNV) that potentially disrupt phosphosites. We found DeepSignal is more sensitive in detecting known cancer genes related to signaling transduction than an existing statistical approach [6]. DeepSignal can further discover new perturbed pathways related to cancer including CTNNB1 pathway in UCEC, PTEN pathway in GBM and SMAD4 pathway in LUAD.

Acknowledgments. This work was supported in part by the NSF CAREER Award, the Sloan Research Fellowship, and the PhRMA Foundation Award in Informatics.

References

1. Xue, Y., et al.: Gps 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol. Cell. Proteomics **7**, 1598–1608 (2008)
2. Miller, M.L., et al.: Linear motif atlas for phosphorylation-dependent signaling. Sci. Signal **1**, ra2 (2008)
3. Li, L., et al.: Prediction of phosphotyrosine signaling networks using a scoring matrix-assisted ligand identification approach. Nucleic Acids Res. **36**, 3263–3273 (2008)
4. Kundu, K., Costa, F., Huber, M., Reth, M., Backofen, R.: Semi-supervised prediction of sh2-peptide interactions from imbalanced high-throughput data. PloS one **8**, e62732 (2013)
5. Zhang, Q.C., Petrey, D., Garzón, J.I., Deng, L., Honig, B.: Preppi: a structure-informed database of protein-protein interactions. Nucleic Acids Res. **41**, D828–D833 (2012)
6. AlQuraishi, M., Koytiger, G., Jenney, A., MacBeath, G., Sorger, P.K.: A multi-scale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks. Nat. Genet. **46**, 1363–1371 (2014)

Integrative Inference of Subclonal Tumour Evolution from Single-Cell and Bulk Sequencing Data

Salem Malikic[1], Katharina Jahn[2,3], Jack Kuipers[2,3], S. Cenk Sahinalp[4(✉)], and Niko Beerenwinkel[2,3(✉)]

[1] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
[2] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
niko.beerenwinkel@bsse.ethz.ch
[3] SIB Swiss Institute of Bioinformatics, Basel, Switzerland
[4] Department of Computer Science, Indiana University, Bloomington, IN, USA
cenksahi@indiana.edu

Cancer is a genetic disease that develops through a branched evolutionary process. It is characterised by the emergence of genetically distinct subclones through the random acquisition of mutations at the level of single-cells and shifting prevalences at the subclone level through selective advantages purveyed by driver mutations. This interplay creates complex mixtures of tumour cell populations which exhibit different susceptibility to targeted cancer therapies and are suspected to be the cause of treatment failure. Therefore it is of great interest to obtain a better understanding of the evolutionary histories of individual tumours and their subclonal composition.

Most of the current data on tumour genetics stems from short read bulk sequencing data. While this type of data is characterised by low sequencing noise and cost, it consists of aggregate measurements across a large number of cells. It is therefore of limited use for the accurate detection of the distinct cellular populations present in a tumour and the unambiguous inference of their evolutionary relationships. Single-cell DNA sequencing instead provides data of the highest resolution for studying intra-tumour heterogeneity and evolution, but is characterised by higher sequencing costs and elevated noise rates.

As the strengths and weaknesses of bulk and single-cell sequencing data are to a large extent complimentary with respect to phylogeny inference, using both data types for a joint inference should improve our understanding of subclonal tumour evolution over using each type of data alone. In this work, we develop B-SCITE, the first computational approach that infers trees of tumour evolution from combined bulk and single-cell sequencing data. B-SCITE employs an MCMC search scheme to find the mutation tree that maximizes the joint likelihood of both data types. The model accounts for typical sequencing biases and artifacts, including the variability in depth of coverage among different bulk sequencing datasets and the contamination of single-cell data by doublets. Using

S. Malikic, K. Jahn, and J. Kuipers — Equal contributors.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 269–270, 2018.
https://doi.org/10.1007/978-3-319-89929-9

a comprehensive set of simulated data, we show that B-SCITE systematically outperforms existing methods with respect to tree reconstruction accuracy and subclone identification. High-fidelity reconstructions are obtained even with a modest number of single cells, suggesting that combined bulk and single-cell data may be a competitive strategy for tumor phylogeny reconstruction. On real data, we show that B-SCITE provides more realistic mutation histories compared to the results reported in previous studies or obtained by existing methods.

Mantis: A Fast, Small, and Exact Large-Scale Sequence-Search Index

Prashant Pandey[1(✉)], Fatemeh Almodaresi[1], Michael A. Bender[1], Michael Ferdman[1], Rob Johnson[1,2], and Rob Patro[1]

[1] Computer Science Department, Stony Brook University, Stony Brook, USA
{ppandey,falmodaresit,bender,mferdman,rob.patro}@cs.stonybrook.edu,
robj@vmware.com
[2] VMware Research, Palo Alto, USA

The ability to issue sequence-level searches over publicly available databases of assembled genomes and known proteins has played an instrumental role in many studies in the field of genomics, and has made BLAST [2] and its variants some of the most widely-used tools in all of science. However, until recently, tools for searches over genomic data were restricted to reference sequences. As a result, the vast majority of publicly-available sequencing data (e.g., the data deposited in the SRA [3]) has been difficult to search because it exists in the form of raw, unassembled sequencing reads.

Recently, Solomon and Kingsford introduced the sequence Bloom tree (SBT) [8] for performing searches over thousands of sequencing experiments. This seminal work introduced both a formulation of this problem, and the initial steps toward a solution. The space and query time of the SBT structure has been further improved by Solomon and Kingsford [9] and Sun et al. [10].

Sequence Bloom trees repurpose Bloom filters to index large sets of raw sequencing data probabilistically and, as a result, they are forced to cope with Bloom filters' limitations. For example, the SBT needs to merge Bloom filters, but Bloom filters must be the same size to be merged, and they cannot be resized. Consequently, SBTs use Bloom filters of the same size to represent sets of widely varying cardinalities. As a result, most of the Bloom filters in the SBT are sub-optimally tuned and inefficient in their use of space. (SBTs partially mitigate this issue by compressing their Bloom filters using an off-the-shelf compressor.)

We introduce Mantis, a space-efficient data structure that can be used to index thousands of raw-read experiments and facilitate large-scale sequence searches on those experiments. Mantis uses counting quotient filters [5] instead of Bloom filters, enabling rapid index builds and queries, small indexes, and *exact* results, i.e., no false positives or negatives. Furthermore, Mantis is also a colored De Bruijn graph (cDBG) representation, and supports the same fast de Bruijn graph traversals as Squeakr [4], and hence may be useful for topological analyses such as computing the length of the query covered in each experiment (rather than just the fraction of k-mers present).

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 271–273, 2018.
https://doi.org/10.1007/978-3-319-89929-9

Mantis has several advantages over prior work:

- Mantis is *exact*. A query for a set Q of k-mers and threshold θ returns exactly those data sets containing at least fraction θ of the k-mers in Q. There are no false positives or false negatives. In contrast, we show that SBT-based systems exhibit only 57–67% precision, meaning that many of the results returned for a given query are, in fact, false positives.
- Mantis supports much faster queries than existing SBT-based systems. In our experiments, queries in Mantis ran up to 100× faster than when using an (in RAM) SSBT.
- Mantis supports much faster index construction. For example, we were able to build the Mantis index on 2,652 data sets in 16 hours and 35 min. SSBT reported 97 hours to construct an index on the same collection of data sets.
- Mantis uses less storage than SBT-based systems. For example, the Mantis index over the 2,652 experiments used for evaluation is 20% smaller than the compressed SSBT index.
- Mantis returns, for each experiment containing at least 1 k-mer from the query, the number of query k-mers present in this experiment. Thus, the full spectrum of relevant experiments can be analyzed. While these results can be post-processed to filter out those not satisfying a θ-query, we believe the Mantis output is more useful, as one can analyze which experiments were close to achieving the θ threshold, and can examine if a natural filtering "cutoff" exists.

Mantis builds on Squeakr, a k-mer counter based on the counting quotient filter (CQF). Prior work has shown how CQFs can be used to improve performance and simplify the design of k-mer-counting tools [4] and de Bruijn graph representations [6].

In a similar spirit, Mantis uses the CQF to create a simple space- and time-efficient index for searching for sequences in large collections of experiments. Mantis is based on cDBGs. The "color" associated with each k-mer in a cDBG is the set of experiments in which that k-mer occurs (similar to Rainbowfish [1]). We use an exact CQF to store a table mapping each k-mer to a color ID, and another table mapping color IDs to the actual set of experiments containing that k-mer. Mantis uses an off-the-shelf compressor [7] to store the bit vectors representing each set of experiments.

Mantis takes as input the collection of CQFs representing each data set, and outputs the search index. Construction is efficient because it can use sequential I/O to read the input and write the output CQFs. Similarly, queries for the color of a single k-mer are efficient since they require only two table lookups.

Mantis is available at https://github.com/splatlab/mantis.

References

1. Almodaresi, F., Pandey, P., Patro, R.: Rainbowfish: A Succinct Colored de Bruijn Graph Representation. In WABI, volume 88, pages 18:1–18:15, 2017
2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W.: Basic local alignment search tool. J. Mol. Biol. **215**, 403–410 (1990)
3. Kodama, Y., Shumway, M., Leinonen, R.: The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. **40**(D1), D54–D56 (2011)
4. Prashant Pandey, Michael A Bender, Rob Johnson, and Rob Patro. Squeakr: An Exact and Approximate k-mer Counting System. Bioinformatics, page btx636, 2017
5. Prashant Pandey, Michael A. Bender, Rob Johnson, and Robert Patro. A General-Purpose Counting Filter: Making Every Bit Count. In SIGMOD, pages 775–787, 2017
6. Prashant Pandey, Michael A. Bender, Rob Johnson, and Robert Patro. deBGR: an efficient and near-exact representation of the weighted de Bruijn graph. Bioinformatics, 33(14), 2017
7. Rajeev Raman, Venkatesh Raman, and S. Srinivasa Rao. Succinct indexable dictionaries with applications to encoding k-ary trees and multisets. In SODA, pages 233–242, 2002
8. Solomon, B., Kingsford, C.: Fast search of thousands of short-read sequencing experiments. Nat. Biotechnol. **34**(3), 300–302 (2016)
9. Solomon, B., Kingsford, C.: Improved Search of Large Transcriptomic Sequencing Databases Using Split Sequence Bloom Trees. In: Sahinalp, S.C. (ed.) RECOMB 2017. LNCS, vol. 10229, pp. 257–271. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56970-3_16
10. Sun, C., Harris, R.S., Chikhi, R., Medvedev, P.: AllSome Sequence Bloom Trees. In: Sahinalp, S.C. (ed.) RECOMB 2017. LNCS, vol. 10229, pp. 272–286. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56970-3_17

Tensor Composition Analysis Detects Cell-Type Specific Associations in Epigenetic Studies

Elior Rahmani[1(✉)], Regev Schweiger[2], Saharon Rosset[3], Sriram Sankararaman[1], and Eran Halperin[1,4,5]

[1] Department of Computer Science, UCLA, Los Angeles, CA, USA
elior.rahmani@gmail.com
[2] Blavatnik School of Computer Science, Tel-Aviv University, Tel Aviv, Israel
[3] Department of Statistics, Tel Aviv University, Tel Aviv, Israel
[4] Department of Human Genetics, UCLA, Los Angeles, CA, USA
[5] Department of Anesthesiology and Perioperative Medicine, UCLA, Los Angeles, CA, USA
ehalperin@cs.ucla.edu

Abstract. Identifying cell-type specific associations of genes with disease and mapping known associations to particular cell types is a key in understanding disease etiology. While developments in technologies for profiling genomic features such as gene expression and DNA methylation have led to the availability of large-scale tissue-specific genomic data, prohibitive costs drastically restrict collection of cell-type specific genomic data. This, in turn, limits the identification of disease-related genes and cell types. It is therefore desired to develop new approaches for detecting cell-type specific associations between phenotypes and tissue-specific genomic data.

We suggest a new matrix factorization formulation, which allows us to deconvolve a two-dimensional input (observations by features) into a three-dimensional output. Traditional matrix factorization formulations essentially take as an input a multiple-source heterogeneous matrix of observations and output a matrix of source-specific weights and a matrix of source-specific features. We generalize this approach by assuming that source-specific features are unique for each observation rather than shared across all observations, and we propose Tensor Composition Analysis (TCA), a method for estimating observation- and source-specific values based on the model.

We apply our model in the context of epigenetic association studies, where DNA methylation data measured from a heterogeneous tissue are often used, and we show that TCA allows us to extract cell-type specific methylation levels from two dimensional tissue-specific methylation data. We further derive a statistical test for detecting cell-type specific effects of methylation on phenotypes based on the TCA model, and using a simulation study we demonstrate its potentials and limitations. Finally, using five large whole-blood methylation datasets, we demonstrate that our model allows the detection of novel replicating cell-type specific associations without collecting cost prohibitive cell-type specific data, thus

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 274–275, 2018.
https://doi.org/10.1007/978-3-319-89929-9

suggesting an exciting new opportunity to unveil more of the hidden signals in genomic association studies with potential design implications for future data collection efforts.

Assembly-Free and Alignment-Free Sample Identification Using Genome Skims

Shahab Sarmashghi[1(✉)], Kristine Bohmann[2,3], M. Thomas P. Gilbert[2,4], Vineet Bafna[5], and Siavash Mirarab[1]

[1] Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA
ssarmash@ucsd.edu
[2] Evolutionary Genomics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
[3] School of Biological Sciences, University of East Anglia, Norwich, Norfolk, UK
[4] Norwegian University of Science and Technology, University Museum, 7491 Trondheim, Norway
[5] Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA

Extended abstract

The ability to quickly and inexpensively describe the taxonomic diversity in an environment is critical in this era of rapid climate and biodiversity changes. The currently preferred molecular technique, barcoding, is low-cost and widely used, but has drawbacks. As sequencing costs continue to fall, an alternative approach based on *genome-skimming* has been proposed [1, 2]. This approach first applies low-pass (100 Mb – several Gb per sample) sequencing to voucher and/or query samples and then recovers marker genes and/or organelle genomes computationally. In contrast, we suggest the use of the unassembled sequence data for taxonomic identification using an alignment-free approach based on the k-mer decomposition of the sequencing reads. Specifically, we first estimate the average sequencing depth and error rate for each genome skim, by comparing our derived theoretical distribution of k-mers' multiplicity and the histogram of k-mer counts computed using Jellyfish [3]. The genome length is also estimated from the average sequencing depth accordingly. Then, the similarity of two genome skims is measured by the Jaccard index between their corresponding k-mer collections. Finally, the hamming distance between genomes is estimated from the Jaccard index, using the following formula obtained by modeling the impact of low sequencing coverage, sequencing error, and differing genome lengths on the similarity of genome skims:

$$D = 1 - \left(\frac{2(\zeta_1 L_1 + \zeta_2 L_2)J}{\eta_1 \eta_2 (L_1 + L_2)(1 + J)} \right)^{1/k} .$$

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 276–277, 2018.
https://doi.org/10.1007/978-3-319-89929-9

In this equation, when coverage is low, we use all k-mers and set:

$$\eta_i = 1 - e^{-c_i(1-k/\ell)(1-\epsilon_i)^k}\ , \quad \zeta_i = \eta_i + c_i(1-k/\ell)(1-(1-\epsilon_i)^k)\ .$$

For higher coverages, we remove k-mers with multiplicity below a threshold m, and set:

$$\zeta_i = \eta_i = 1 - \sum_{t=0}^{m-1} \frac{(c_i(1-k/\ell)(1-\epsilon_i)^k)^t}{t!} e^{-c_i(1-k/\ell)(1-\epsilon_i)^k}\ .$$

In these equations, k and ℓ are k-mer and read length, respectively, and c_i, ϵ_i, and L_i are substituted from the estimates of coverage, error rate, and genome length for each genome skim. The Jaccard index between two genome skims, J, is computed by Mash [4] efficiently using a hashing technique.

We have tested our tool, Skmer, on genome skims simulated from assemblies of 90 species from two genera of insects (Anopheles and Drosophila) and across the avian tree of life. We test the accuracy of the distances computed by Skmer, and subsequently use the distances to find the exact/closest match to a query sample in a reference set of genome skims. Comparing to the other k-mer based tools, Skmer shows excellent performance in our simulation studies, especially when the coverage is below 4X [5].

Skmer makes the assembly-free approach to genome-skimming a viable alternative to the traditional barcoding. The software is made publicly available on Github (https://github.com/shahab-sarmashghi/Skmer.git).

References

1. Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C., Liston, A.: Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am. J. Bot. **99**(2), 349–364 (2012)
2. Coissac, E., Hollingsworth, P.M., Lavergne, S., Taberlet, P.: From barcodes to genomes: extending the concept of dna barcoding. Mol. Ecol. **25**(7), 1423–1428 (2016)
3. Marçais, G., Kingsford, C.: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**(6), 764–770 (2011)
4. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M.: Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. **17**(1), 132 (2016)
5. Sarmashghi, S., Bohmann, K., Gilbert, M.T.P., Bafna, V., Mirarab, S.: Assembly-free and alignment-free sample identification using genome skims (2017). bioRxiv 230409

Efficient Algorithms to Discover Alterations with Complementary Functional Association in Cancer

Rebecca Sarto Basso[1], Dorit S. Hochbaum[1], and Fabio Vandin[2,3(✉)]

[1] University of California at Berkeley, Berkeley, USA
rebeccasarto@berkeley.edu, hochbaum@ieor.berkeley.edu
[2] University of Padova, Padova, Italy
fabio.vandin@unipd.it
[3] Brown University, Providence, USA

Introduction. Recent advances in sequencing technologies now allow to assay the entire complement of somatic alterations in large tumour cohorts [5]. Several computational methods have been recently designed to identify *driver* alterations, associated to the disease, and to distinguish them from *passenger* alterations not related with the disease. The identification of driver alterations is complicated by the extensive *intertumour heterogeneity*, with large (100–1000's) and different collections of alterations being present in tumours from different patients and no two tumours having the same collection of alterations [6, 7]. One of the reasons for such heterogeneity is that driver alterations target cancer *pathways*, groups of interacting genes performing given functions in the cell and whose alteration is required to develop the disease [2, 7]. One of the main remaining challenges is the identification of alterations with functional impact [3].

Several methods for the *de novo* discovery of mutated cancer pathways have leveraged the *mutual exclusivity* of cancer alterations, with cancer pathways displaying at most one alteration for each patient [3, 7]. The mutual exclusivity property is due to the complementarity of genes in the same pathway, with alterations in different members of a pathway resulting in a similar impact at the functional level. An additional source of information that can be used to identify genes with complementary functions are quantitative measures for each samples such as functional profiles, obtained for example by genomic or chemical perturbations [1]. The employment of such quantitative measurements is crucial to identify meaningful complementary alterations since one can expect mutual exclusivity to reflect in functional properties of altered samples which are specific to the altered samples.

Methods and Results. We study the problem of finding sets of alterations with complementary functional associations using alteration data and a quantitative (functional) target measure from a collection of cancer samples. We provide a rigorous combinatorial formulation for the problem and prove that the associated computational problem is NP-hard. We develop two efficient algorithms, a greedy algorithm and an ILP-based algorithm to identify the set of k genes with the highest association with a target and prove rigorous guarantees in the quality of their solutions.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 278–279, 2018.
https://doi.org/10.1007/978-3-319-89929-9

Our algorithms are implemented in our tool fUNctional Complementarity of alteratiOns discoVERy (UNCOVER)[1]. We compared UNCOVER with REVEALER [4], a recently developed greedy algorithm to identify mutually exclusive sets of alterations associated with functional phenotypes. Considering four cancer datasets from [4], we compared the solutions obtained by our algorithms with the solutions from REVEALER in terms of the *information coefficient* (IC), the target association score used in [4] as a quality of the solution. Surprisingly, in two out of four datasets our methods, which do not consider the IC score, identify solutions with IC score *higher* (by at least 5%) than the solutions reported by REVEALER, while for the other two datasets the IC score is very similar. These results show that UNCOVER identifies better solutions than REVEALER when evaluated using our objective function *and* also when evaluated according to the objective function of REVEALER.

In addition, UNCOVER has a running time that is on average two orders of magnitude smaller than required by REVEALER. The efficiency of UNCOVER enables the analysis of a large number of targets. We have run UNCOVER on a dataset with thousands of functional targets and tens of thousands alterations from the Achilles project dataset[2] and the Cancer Cell Line Encyclopedia (CCLE). While running UNCOVER (including preprocessing) on the entire dataset required 24 h, based on the runtime required on the instances reported in [4] running REVEALER on this dataset would have required about 5 months of compute time. On such large dataset, UNCOVER identifies several statistically significant associations between target values and mutually exclusive alterations in genes sets.

Acknowledgement. This work is supported, in part, by NSF grant IIS-124758 and by the University of Padova grants SID2017 and PROACTIVE2017. This work was done in part while FV was visiting the Simons Institute for the Theory of Computing, supported by the Simons Foundation.

References

1. Cowley, et al.: Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Sci Data (2014)
2. Creixell, et al.: Pathway and network analysis of cancer genomes. Nat. Met. (2015)
3. Garraway and Lander: Lessons from the cancer genome. Cell (2013)
4. Kim, et al.: Characterizing genomic alterations in cancer by complementary functional associations. Nat. Biotech. (2016)
5. TCGA Research Network: Integrated genomic characterization of pancreatic ductal adenocarcinoma. Cancer Cell (2017)
6. Vandin: Computational methods for characterizing cancer mutational heterogeneity. Frontiers in genetics (2017)
7. Vogelstein, et al.: Cancer genome landscapes. Science (2013)

[1] https://github.com/VandinLab/UNCOVER.
[2] https://portals.broadinstitute.org/achilles.

Latent Variable Model for Aligning Barcoded Short-Reads Improves Downstream Analyses

Ariya Shajii[1], Ibrahim Numanagić[1,2], and Bonnie Berger[1,2(✉)]

[1] Computer Science and AI Lab, MIT, Cambridge, MA, USA
bab@mit.edu
[2] Department of Mathematics, MIT, Cambridge, MA, USA

Background: Barcoded read sequencing allows short-reads to carry long-range information by virtue of read "barcodes", and has several advantages (including significantly reduced cost and lower error rates) over long-read sequencing. Here we introduce a two-tiered statistical binning approach, EMerAld—or EMA for short—to barcoded read sequence alignment, an essential component of any barcoded sequencing pipeline, and as a result improve downstream genotyping and phasing. Our method enables the probabilistic placement of reads between different read clouds [1], and also in a single cloud that spans homologous elements. The two tiers consist of: (i) a novel latent variable model to probabilistically assign reads to possible source fragments; and (ii) newly exploiting expected read coverage (read density) to resolve the difficult case of multiple repetitive alignments of reads within a single read cloud. These ambiguous alignments account for a large fraction of the rare variants that currently cannot be resolved and are of great interest to biologists [2].

Methods: Current linked-read alignment methods first perform a standard all-mapping, then partition the resulting alignments into groups of nearby reads with a common barcode called "read clouds". Reads are then assigned to one of their possible clouds by optimizing a global score function that takes into account edit distance, mate pairs, read clouds, etc. Our two main conceptual advances are as follows. Intuitively, rather than assigning each read to just one of its possible alignments at any given time, we make use of probabilistic assignments of reads to clouds and employ a latent variable model to determine final alignment probabilities; thereby, we select the most likely cloud (and thus alignment) for each read. During the cloud alignment process, we also utilize a disjoint-set data structure over read clouds to normalize alignment probabilities in a physically sensible way. Once reads are assigned to clouds, we propose a different statistical binning optimization approach to better handle the ubiquitous repetitive regions of the genome. Whereas currently-used methods simply pick the lowest edit distance alignment of a read in a given cloud, we instead optimize a combination of edit distance and "read density", which takes into account the read density distribution over fragments. This two-tiered process can be interpreted

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 280–282, 2018.
https://doi.org/10.1007/978-3-319-89929-9

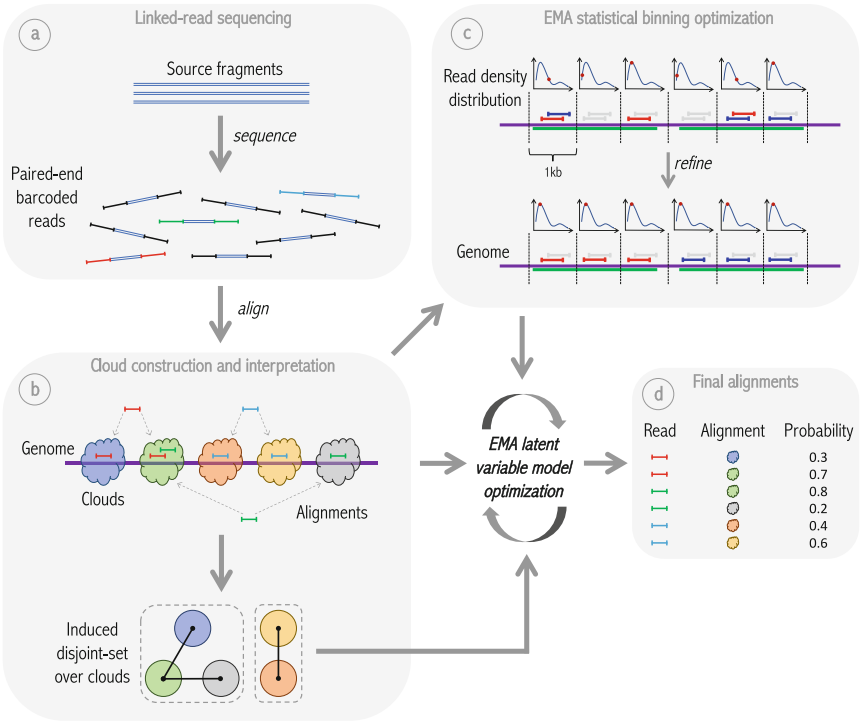

Fig. 1. Overview of EMA pipeline. **(a)** Idealized model of linked-read sequencing, wherein some number of unknown source fragments in a single droplet are sheared, barcoded and sequenced to produce linked-reads. **(b)** EMA's "read clouds" are constructed by grouping nearby-mapping reads sharing the same barcode; these clouds represent possible source fragments. EMA then partitions the clouds into a disjoint-set induced by the alignments, where two clouds are connected if there is a read aligning to both; connected components in this disjoint-set (enclosed by dashed boxes) correspond to alternate possibilities for the *same* unknown source fragment. EMA's latent variable model optimization is subsequently applied to each of these connected components individually. **(c)** EMA applies a novel statistical binning optimization algorithm to clouds containing multiple alignments of the same read to pick out the most likely alignment, by optimizing a combination of alignment edit distances and read densities within the cloud. In the figure, the green regions of the genome are homologous, thereby resulting in multi-mappings within a single cloud. **(d)** While the statistical binning optimization operates within a single cloud, EMA's latent variable model optimization determines the best alignment of a given read between different clouds, and produces not only the final alignment for each read, but also interpretable alignment *probabilities*.

as statistical binning first in assigning reads to clouds and then within clouds. The EMA pipeline is shown in Fig. 1.

Results: EMA is much faster and less memory intensive compared to other tools. EMA's overhead over the initial run of an all-mapper is virtually negligible,

and EMA is at least 1.5× faster than Lariat (the current 10x alignment tool [1]), which translates into days faster for the user. In addition, we show that genotypes called from EMA's alignments contain over 30% fewer false positives than those called from Lariat's, with a fewer number of false negatives, on 10x WGS datasets of NA12878 and NA24385, as compared to NIST GIAB gold standard variant calls. We also demonstrate that EMA's alignments improve phasing performance over Lariat's in both NA12878 and NA24385, producing fewer switch/mismatch errors and larger phased blocks on average.

Moreover, we demonstrate that EMA is able to effectively resolve alignments in regions containing nearby homologous elements—a particularly challenging problem in read mapping—through the introduction of our novel statistical binning optimization framework, which enables us to find variants in the pharmacogenomically important CYP2D region that go undetected when using Lariat or BWA. This enhanced capability addresses one of the major weaknesses of linked-read sequencing as compared to long-read sequencing, where only a relatively small subset of the original source fragment is observed—and more specifically, that the order of reads within the fragment is not known—making it difficult to produce accurate alignments if the fragment spans homologous elements.

Discussion: Our advance is a general framework applicable to many barcoded sequencing problems. It is likely to be of interest to any developers, and even users, of barcoded or linked-read sequencing technologies that come along. We highlight that 10x sequencing is just an instance of general "barcoded read sequencing", and other technologies that make use of the same paradigm already exist and are likely to emerge in the future, given its numerous advantages over long-read sequencing. Several technologies already employ barcoded sequencing in addition to 10x Genomics', such as Illumina's TruSeq SLR platform (formerly Moleculo), and Complete Genomics' Long Fragment technology. Our framework should apply to these (and similar) technologies as well. Due to their substantial improvements over existing methods for aligning and interpreting linked-read data, the algorithms employed by EMA are likely to be a fundamental component of read cloud-based methods in the future.

Acknowledgements. We thank Chris Whelan, Chad Nusbaum, Eric Banks, as well as the rest of the SV Group from the Broad Institute for providing us with data samples and many valuable suggestions. Also, we thank Jian Peng and Lillian Zhang for their helpful suggestions.

Funding A.S., I.N. and B.B. are partially funded by NIH grant GM108348.

References

1. Bishara, A., et al.: Read clouds uncover variation in complex regions of the human genome. Genome Res **25**(10), 1570–1580 (2015)
2. Sekar, A., et al.: Schizophrenia risk from complex variation of complement component 4. Nature **530**, 177 (2016)

ModulOmics: Integrating Multi-Omics Data to Identify Cancer Driver Modules

Dana Silverbush[1(✉)], Simona Cristea[2,3,4(✉)], Gali Yanovich[5], Tamar Geiger[5], Niko Beerenwinkel[6,7], and Roded Sharan[1]

[1] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel
dsilverb@broadinstitute.org, roded@post.tau.ac.il
[2] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA
[3] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[4] Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA
scristea@jimmy.harvard.edu
[5] Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
[6] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
niko.beerenwinkel@bsse.ethz.ch
[7] Swiss Institute of Bioinformatics, Basel, Switzerland

Introduction: Recent rapid advancements in sequencing technologies allowed the collection of DNA, RNA, and protein data from tens of thousands of cancer patients. Mathematical and computational tools are used to analyze these complex data sets, aiming to reveal mechanistic and predictive insights into tumor treatment and progression. Key to achieving these goals is finding molecular alterations that drive tumorigenesis, or drivers, such as single nucleotide variants (SNVs), copy number alterations (CNAs), changes in the transcriptional activity of genes, or changes in protein concentration. Groups of such functionally connected genetic alterations, also termed cancer driver modules or pathways, activate mechanisms that gradually contribute to triggering the hallmarks of cancer, conferring fitness advantages to the tumors. The identification of such driver modules is an important challenge in the field of cancer genomics, since clinically targeting driver pathways can improve patient treatment. Nevertheless, most of the existing computational tools to address this problem use primarily somatic mutations, not fully exploiting additional data types. Here, we describe ModulOmics, a method to *de novo* identify cancer driver modules by integrating multiple sources of biological information (protein-protein interactions, mutual exclusivity of mutations or copy number alterations, transcriptional co-regulation, and RNA co-expression) into a single probabilistic model.

Methods: Given a set $G = \{G_1, \ldots, G_n\}$ of genes and a collection $M = \{M_1, \ldots M_m\}$ of models for different data types, we introduce S_G,

D. Silverbush, S. Cristea, N. Beerenwinkel, and R. Sharan — equal contribution.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 283–284, 2018.
https://doi.org/10.1007/978-3-319-89929-9

the ModulOmics probabilistic score of the set G, reflecting how likely are the genes in G to be functionally connected. S_G is computed as the mean of m probabilistic scores $P(G \mid M_k)$, each representing the degree of functional connectivity of the set G, under a different model:

$$S_G = \frac{1}{m} \sum_{k=1}^{m} P(G \mid M_k) \tag{1}$$

Here, we consider four models, as follows: M_1 computes the connectivity of the genes in G based on their proximity in the protein-protein interaction (PPI) network, M_2 estimates the degree of mutual exclusivity among DNA alterations of the genes in G across the patient cohort, M_3 assesses the co-regulation of the genes in G on the basis of their shared transcriptional regulators that are active in the patient cohort, and M_4 evaluates the transcriptional connectivity of the genes in G based on their coexpression profiles. The goal of ModulOmics is to identify groups that maximize the global score in Eq. 1. As the number of candidate groups grows exponentially with maximal group size, we use a heuristic two-step optimization procedure. The optimization routine first performs an approximation of the exact scores of the set G under each of the four models M_k, by decomposing them into pairwise scores and using integer linear programming (ILP) to find good initial solutions. The initial solutions are further refined via stochastic search starting from these initial solutions and using the global score.

Results: Using ModulOmics, we accurately identify known cancer driver genes and pathways in three large-scale TCGA datasets of breast cancer, glioblastoma (GBM) and ovarian cancer, outperforming state-of-the-art methods for module detection. Notably, in breast cancer subtypes, the highest scoring modules reliably separate cancerous from normal tissues in an independent patient cohort. Focusing on individual subtypes, the modules of Her2 and Basal are enriched with Gene Ontology (GO) terms related to cell proliferation, reflecting their more aggressive nature. Driver modules in triple negative (TN) samples capture the accumulation of down-regulated tumor suppressors such as *TP53*, *BRCA1*, *RB1* and *PTEN*, a pattern also supported by reverse phase protein array (RPPA) data. The highest scoring modules in Luminal A suggest two potential functionalities of *PTEN*: a canonical one as part of the PI3K pathway, and a non-canonical one as a regulator of cell proliferation. ModulOmics is freely available in two forms, as an open-source R code for the identification of cancer driver modules from a cohort of cancer samples (https://github.com/danasilv/ModulOmics), and as a webserver for the evaluation of any set of genes of interest using the TCGA data processed in this study (http://anat.cs.tau.ac.il/ModulOmicsServer/).

SCIΦ: Single-Cell Mutation Identification via Phylogenetic Inference

Jochen Singer[1,2], Jack Kuipers[1,2], Katharina Jahn[1,2], and Niko Beerenwinkel[1,2(✉)]

[1] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
niko.beerenwinkel@bsse.ethz.ch
[2] SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Abstract. Understanding the evolution of cancer is important for the development of appropriate cancer therapies. The task is challenging because tumors evolve as heterogeneous cell populations with an unknown number of genetically distinct subclones of varying frequencies. Conventional approaches based on bulk sequencing are limited in addressing this challenge as clones cannot be observed directly. Single-cell sequencing holds the promise of resolving the heterogeneity of tumors. However, this advantage comes at the cost of elevated noise due to the limited amount of DNA material present in a cell and the extensive DNA amplification required prior to sequencing.

Here, we present SCIΦ, the first single-cell-specific variant caller that combines single-cell genotyping with reconstruction of the cell lineage tree. SCIΦ leverages the fact that the somatic cells of an organism are related via a phylogenetic tree where mutations are propagated along tree branches. Our inference scheme starts with an initial identification of possible mutation loci and then performs joint phylogenetic inference and variant calling via posterior sampling.

In a first step, likely mutated loci are identified using the posterior probability of observing at least one mutated cell at a specific locus. In order to do so, SCIΦ models the nucleotide counts using a beta-binomial distribution. This is especially useful in the single-cell setting, since the beta-binomial distribution can be described as a Pólya urn model, which in turn is a very close approximation of the multiple displacement amplification commonly used to amplify the genomic material of a single-cell.

In a second step, the identified loci are used to infer the tumor phylogeny. Here, we account for dropout events by modeling the likelihood of observing a mutation in a cell as a weighted mixture of the likelihoods of homozygous reference genotype, heterozygous genotype, and homozygous alternative genotype. Our model to infer tumor phylogeny consists of three parts: the genealogical tree, the mutation attachments to edges, and the parameters of the model. Because the tree search space grows superexponentially in the number of cells, we employ a Markov Chain

J. Singer, J. Kuipers — These authors contributed equally.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 285–286, 2018.
https://doi.org/10.1007/978-3-319-89929-9

Monte Carlo scheme to traverse through the tree space with mutation assignment and learn the parameters of the model.

Using the relationship between cells, we are able to reliably call mutations in each single-cell even in experiments with high dropout rates and missing data. We show that SCIΦ outperforms existing methods on simulated data and apply it to different real-world datasets. Availability: https://github.com/cbg-ethz/SCIPhI

AptaBlocks: Accelerating the Design of RNA-Based Drug Delivery Systems

Yijie Wang[1], Jan Hoinka[1], Piotr Swiderski[2], 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 91010, USA

Extended Abstract

Synthetic RNA molecules are increasingly used to alter cellular functions [1–4]. These successful applications indicate that RNA-based therapeutics might be able to target currently undruggable genes [5, 6]. However, to achieve this promise, an effective method for delivering therapeutic RNAs into specific cells is required. Recently, RNA aptamers emerged as promising delivery agents due to their ability of binding specific cell receptors [7, 8]. Crucially, these aptamers can frequently be internalized into the cells expressing these receptors on their surfaces. This property is leveraged in aptamer based drug delivery systems by combining such receptor-specific aptamers with a therapeutic "cargo" such that the aptamer facilitates the internalization of the cargo into the cell [9–11]. The advancement of this technology however is contingent on an efficient method to produce stable molecular complexes that include specific aptamers and cargoes. A recently proposed experimental procedure for obtaining such complexes relies on conjugating the aptamer and the cargo with complementary RNA strands so that when such modified molecules are incubated together, the complementary RNA strands hybridize to form a double-stranded "sticky bridge" connecting the aptamer with its cargo [12, 13]. However, designing appropriate sticky bridge sequences guaranteeing the formation and stability of the complex while simultaneously not interfering with the aptamer or the cargo as well as not causing spurious aggregation of the molecules during incubation has proven highly challenging.

To fill this gap, we developed AptaBlocks, a computational method to design sticky bridges to connect RNA-based molecules (blocks). Accounting for the three-step procedure [12, 13], we formulate the sticky bridge sequence design as an optimization problem utilizing an objective function which reflects the biophysical characteristics of the assembly process. Specifically, we designed the objective function considering the equilibrium probabilities of the target structures over all possible structures of the aptamer-stick and cargo-stick, the probability of the interaction between the aptamer-stick and cargo-stick at equilibrium, the hybridization energy between the sticky bridge sequences, and additional

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 287–288, 2018.
https://doi.org/10.1007/978-3-319-89929-9

sequence constraints including but not limited to the GC content. We further provide a simulated annealing algorithm that enables efficient estimation of the corresponding combinatorial optimization problem. The effectiveness of the algorithm has been verified computationally and experimentally. AptaBlocks can be used in a variety of experimental settings and its preliminary version has already been leveraged to design an aptamer based delivery system for a cytotoxic drug targeting Pancreatic ductal adenocarcinoma cells [14]. It is thus expected that AptaBlocks will play a substantial role in accelerating RNA-based drug delivery design.

References

1. Kushwaha, M., et al.: Using RNA as molecular code for programming cellular function. ACS Synth. Biol. **5**(8), 795–809 (2016)
2. Chappell, J., Watters, K.E., Takahashi, M.K.: A renaissance in RNA synthetic biology: new mechanisms, applications and tools for the future. Curr. Opin. Chem. Biol. **28**, 47–56 (2015)
3. Mckeague, M., Wong, R.S., Smolke, C.D.: Opportunities in the design and application of RNA for gene expression control. Nucleic Acids Res. **44**(10), 2987–2999 (2016)
4. Qi, L.S., Arkin, A.P.: A versatile framework for microbial engineering using synthetic non- coding RNAs. Nat. Rev. Microbiol. **12**(5), 341–354 (2014)
5. Ryther, R.C.C., et al.: siRNA therapeutics: big potential from small RNAs. Gene Ther. **17**(1), 5–11 (2005)
6. Chakraborty, C.: Potentiality of small interfering RNAs (siRNA) as recent therapeutic targets for. Curr. Drug Targets **8**(3), 469–482 (2007)
7. Zhou, J., Rossi, J.J.: Cell-specific aptamer-mediated targeted drug delivery. Oligonucleotides **21**(1), 1–10 (2011)
8. Zhang, Y., Hong, H., Cai, W.: Tumor-targeted drug delivery with aptamers. Curr. Med. Chem. **18**(27), 4185–4194 (2011)
9. Mcnamara II, J.O., et al.: Cell type-specific delivery of siRNAs with aptamer siRNA chimeras. Nat. Biotechnol. **24**(8), 1005–1015 (2006)
10. Thiel, K.W., et al.: Delivery of chemo-sensitizing siRNAs to HER2 + -breast cancer cells using RNA aptamers. Nucleic Acids Res. **40**(13), 6319–6337 (2012)
11. Pastor, F., et al.: Induction of tumour immunity by targeted inhibition of nonsense-mediated mRNA decay. Nature **465**(7295), 227–230 (2010)
12. Zhou, J., et al.: Selection, characterization and application of new RNA HIV gp 120 aptamers for facile delivery of Dicer substrate siRNAs into HIV infected cells. Nucleic Acids Res. **37**(9), 3094–3109 (2009)
13. Zhou, J., Rossi, J.: Aptamers as targeted therapeutics: current potential and challenges. Nat. Rev. Drug Discov. **16**(3), 181–202 (2016)
14. Yoon, S., et al.: Aptamer-drug conjugates of active metabolites of nucleo-side analogs and cytotoxic agents inhibit pancreatic tumor cell growth. Mol. Ther.: Nucleic Acid **6**, 80—88 (2017)

A Unifying Framework for Summary Statistic Imputation

Yue Wu[1], Eleazar Eskin[1,2], and Sriram Sankararaman[1,2(✉)]

[1] Department of Computer Science, UCLA, Los Angeles, USA
{eeskin,sriram}@cs.ucla.edu
[2] Department of Human Genetics, UCLA, Los Angeles, USA

Imputation has been widely utilized to aid and interpret the results of Genome-Wide Association Studies(GWAS). Imputation methods, that aim to fill in "data" at untyped SNPs, have emerged as an effective strategy to increase the power of GWAS since the causal variant may not be directly observed or typed in these studies. In the context of GWAS, there are two broad classes of methods to impute association statistics at untyped SNPs. The first class, termed **Two-step imputation**, imputes genotypes at untyped SNPs followed by computing association statistics at the imputed genotypes [1–6]. In practice, the first step of genotype imputation relies on discrete Hidden Markov Models (HMM) [1, 6]. The second class of methods, termed *summary statistic imputation* (**SSI**), directly imputes association statistics at untyped SNPs given the association statistics at the typed SNPs. The joint distribution of association statistics at the typed SNPs and untyped SNPs has been shown to follow a multivariate normal distribution (MVN) [7–9]. **SSI** is appealing as it tends to be computationally efficient while only requiring the summary statistics from a study while the **Two-step imputation** methods require access to individual-level data which can be difficult to obtain in practice.

Current summary-statistic based imputation methods calibrate the imputed statistics using a technique we call *variance re-weighting* (**SSI-VR**). Despite recent progress, the statistical properties of summary statistic imputation methods (including the impact of variance re-weighting) and the connection between the two classes of summary statistic imputation methods has not been adequately understood.

In this paper, we show that the two classes of imputation methods, **Two-step imputation** and **SSI** are asymptotically multivariate normal with small differences in the underlying covariance matrix. Using this asymptotic equivalence, we can understand the effect of the imputation method on the power of the study. Our new method, **SSI**, performs summary statistic imputation without variance re-weighting. The resulting statistics do not then have unit variance as in traditional summary statistic imputation but instead correctly take into account the ambiguity of the imputation process.

We compared the performance of the different imputations methods on the Northern Finland Birth Cohort (NFBC) data set [10] to show that**SSI** increases power over no imputation while SSI-VR can sometimes lead to lower power.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 289–290, 2018.
https://doi.org/10.1007/978-3-319-89929-9

Finally, we compared the results from **SSI**, **SSI-VR** and **Two-step imputation** on the NFBC dataset and show that the resulting statistics are close thereby justifying the theory.

References

1. Browning, S.R., Browning, B.L.: Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. Am. J. Hum. Genet. **81**, 1084–1097 (2007)
2. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., Abecasis, G.R.: Fast and accurate genotype imputation in genome-wide association studies through prephasing. Nat. Genet. **44**(8), 955–959 (2012)
3. Howie, B.N., Donnelly, P., Marchini, J.: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. **5**(6), e1000529 (2009)
4. Li, Y., Willer, C., Sanna, S., Abecasis, G.: Genotype imputation. Annu. Rev. Genomics Hum. Genet. **10**, 387–406 (2009)
5. Li, Y., Willer, C.J., Ding, J., Scheet, P., Abecasis, G.R.: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. **34**(8), 816–834 (2010)
6. Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P.: A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. **39**, 906–913 (2007)
7. Han, B., Kang, H.M., Eskin, E.: Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genet. **5**(4), e1000456 (2009)
8. Kostem, E., Lozano, J.A., Eskin, E.: Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms. Genetics **188**(2), 449–460 (2011)
9. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., Eskin, E.: Identifying causal variants at loci with multiple signals of association. Genetics **198**(2), 497–508 (2014)
10. Sabatti, C., Hartikainen, A.-L., Pouta, A., et al.: Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat. Genet. **41**(1), 35–46 (2009)

Characterizing Protein-DNA Binding Event Subtypes in ChIP-Exo Data

Naomi Yamada, William K. M. Lai, Nina Farrell, B. Franklin Pugh, and Shaun Mahony(✉)

Department of Biochemistry and Molecular Biology, Center for Eukaryotic Gene Regulation, The Pennsylvania State University, University Park, PA 16802, USA
mahony@psu.edu

Introduction: A given regulatory protein may have multiple modes of interaction with the genome; at some sites, it may directly bind cognate DNA motifs, while at others it may bind indirectly via protein-protein interactions with other regulators. Each protein-DNA interaction mode may be associated with distinct sequence motifs, and may also produce distinct patterns in high-resolution protein-DNA binding assays. For example, the ChIP-exo [1] protocol precisely characterizes protein-DNA crosslinking patterns by combining chromatin immunoprecipitation (ChIP) with 5' to 3' exonuclease digestion. Since different regulatory complexes will result in different protein-DNA crosslinking signatures, analysis of ChIP-exo sequencing tag patterns should enable detection of multiple protein-DNA binding modes for a given regulatory protein. However, current ChIP-exo analysis methods either treat all binding events as being of a uniform type, or rely on DNA motifs to cluster binding events into subtypes.

We introduce the ChIP-exo mixture model (ChExMix) to systematically detect multiple protein-DNA interaction modes in a single ChIP-exo experiment. ChExMix discovers and characterizes binding event subtypes in ChIP-exo data by leveraging both sequencing tag enrichment patterns and DNA motifs. ChExMix defines possible binding event subtypes by both clustering observed ChIP-exo tag distribution patterns and performing targeted *de novo* motif discovery around the positions of the predicted binding events. ChExMix then uses an Expectation Maximization learning scheme to probabilistically model the genomic locations and subtype membership of binding events using both ChIP-exo tag locations and DNA sequence information. In analyzing ChIP-exo data, ChExMix offers a more principled and robust approach to characterizing binding subtypes than simply clustering binding events using motifs.

Results: ChExMix uses DNA motif and ChIP-exo tag distribution patterns to accurately estimate multiple binding subtypes within a single ChIP-exo. We demonstrate the ability of ChExMix to estimate binding subtypes and assign binding events to subtypes by creating datasets that computationally mix data from CTCF and FoxA1 ChIP-exo experiments. CTCF and FoxA1 are known to display distinct ChIP-exo tag distribution patterns at their respective binding events. We simulated different representations of each subtype by modulating the relative number of tags drawn from each ChIP-exo experiment. ChExMix detects the two subtypes and accurately assigns subtypes to binding events over a wide range of relative sampling rates from the CTCF and FoxA1 subtypes. In contrast, a motif-driven approach fails to appropriately classify

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 291–292, 2018.
https://doi.org/10.1007/978-3-319-89929-9

many of the FoxA1 subtype binding events. ChExMix performance remains reasonably high when we remove DNA motifs from consideration and assign subtypes using only ChIP-exo tag distribution information. Our results demonstrate that ChExMix enables discovery of unique subtypes within a single ChIP-exo dataset and accurately assigns subtypes to binding events.

To assess ChExMix's ability to characterize binding locations, we compare ChExMix performance in predicting human CTCF and mouse FoxA2 binding event locations to that of seven ChIP-exo analysis methods. ChExMix outperforms other methods by exactly locating the CTCF events at the motif position in 90.2% of the shared CTCF events. Similarly, ChExMix exactly locates the FoxA2 events at the motif position in 67.4% of the shared FoxA2 events. ChExMix binding event predictions also contain instances of the cognate motif at a high rate. These results suggest that ChExMix maintains high accuracy in protein-DNA binding event predictions.

We further demonstrate that ChExMix can characterize biologically relevant binding event subtypes in ER positive breast cancer cells. FoxA1, ERα, and CTCF have previously been shown to co-localize at a subset of genomic loci. However, how these proteins interact with each other and DNA at specific sites remained elusive. In FoxA1 ChIP-exo data, ChExMix identifies subtypes corresponding to ERα and CTCF motifs, and about a half of these subtypes' binding events display ERα and CTCF ChIP-exo enrichment with similar tag distributions. Our results thus suggest that ERα and CTCF may mediate binding of FoxA1 via protein-protein interactions at a subset of the genomic loci where multiple factors are co-bound. These results strongly suggest that ChExMix can discover binding event subtypes representing direct and indirect TF interactions from a single ChIP-exo experiment.

Conclusions: ChExMix provides a principled platform for elucidating diverse protein-DNA interaction modes in a single ChIP-exo experiment by exploiting both ChIP-exo tag enrichment patterns and DNA motifs. Using a fully integrated framework, ChExMix allows simultaneous detection of binding event locations, discovery of binding event subtypes, and assignment of binding events to subtypes. ChExMix enables new forms of insight from a single ChIP-exo experiment, taking analysis towards a fine-grained characterization of distinct protein-DNA binding modes at specific genomic loci. ChExMix is freely available from https://github.com/seqcode/chexmix.

Reference

1. Rhee, H.S., Pugh, B.F.: Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell **147**(6), 1408–1419 (2011)

Continuous-Trait Probabilistic Model for Comparing Multi-species Functional Genomic Data

Yang Yang[1], Quanquan Gu[2], Takayo Sasaki[3], Julianna Crivello[4], Rachel O'Neill[4], David M. Gilbert[3], and Jian Ma[1(✉)]

[1] Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
jianma@cs.cmu.edu

[2] Department of Computer Science, University of Virginia, Charlottesville, USA

[3] Department of Biological Science, Florida State University, Tallahassee, USA

[4] Department of Molecular and Cell Biology, Institute for Systems Genomics, University of Connecticut, Storrs, USA

Multi-species functional genomic data from various high-throughput assays are highly informative for the comparative analysis of gene regulation to better understand the molecular mechanisms of phenotypic diversity between human and other mammalian species. Continuous-trait models, which are key to the modeling of functional genomic signals, are gaining increasing attention in genome-wide comparative genomic studies. However, computational models are currently under-explored to fully capture continuous features in the context of multi-species comparisons. There have been several types of continuous-trait evolutionary models, including Brownian motion and Ornstein-Uhlenbeck (OU) process. However, to the best of our knowledge, there are no existing computational methods available to simultaneously infer heterogeneous continuous-trait evolutionary models along the genome based on functional genomic signals.

In this paper, we develop a new continuous-trait probabilistic model for more accurate state estimation using multi-variate features from cross-species functional genomic signals. We call our model phylogenetic hidden Markov Gaussian processes (Phylo-HMGP). Phylo-HMGP incorporates the evolutionary affinity among multiple species into the hidden Markov model (HMM) for exploiting both temporal dependencies across species in the context of evolution and spatial dependencies along the genome in a continuous-trait model. The goal of the proposed method is to identify heterogeneous cross-species genomic feature patterns more effectively. The Gaussian processes embedded in the HMM are specialized to be multi-variate OU processes or Brownian motion in this study.

Both simulation studies and real data application demonstrate the effectiveness of Phylo-HMGP. Importantly, we applied Phylo-HMGP to analyze a new cross-species DNA replication timing (RT) dataset from the same cell type in five primate species (human, chimpanzee, orangutan, gibbon, and green monkey). We demonstrate that our Phylo-HMGP model enables discovery of genomic regions with distinct evolutionary patterns of RT. We found that regions with

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 293–294, 2018.
https://doi.org/10.1007/978-3-319-89929-9

conserved early RT and conserved late RT exhibit strong correlation with constitutive early RT and constitutive late RT, respectively, defined from human ES cell differentiation. In addition, we found enrichment for specific *cis*-regulatory elements in hominini specific early RT regions.

Taken together, the proposed Phylo-HMGP explores a new integrative framework to utilize continuous-trait evolutionary models with spatial constraints to study genome-wide functional genomic features across species. The new method is also flexible such that varied continuous-trait evolutionary models or assumptions can be incorporated. We believe that Phylo-HMGP provides a generic framework that has the potential to more precisely capture the evolutionary history of regulatory regions based on functional genomic signals across different species.

Deep Learning Reveals Many More Inter-protein Residue-Residue Contacts than Direct Coupling Analysis

Tian-Ming Zhou[1,2], Sheng Wang[3(✉)], and Jinbo Xu[1(✉)]

[1] Toyota Technological Institute at Chicago, Chicago, USA
jinboxu@gmail.com
[2] The Institute for Theoretical Computer Science (ITCS), Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[3] Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

We study how to predict inter-protein residue-residue contacts between a pair of putative interacting proteins, which has been reported useful for the 3D structure modeling of a PPI or protein docking. Direct-coupling analysis (DCA) has been applied to intra-protein and inter-protein contact prediction, but it does not fare well for proteins without many sequence homologs. This is a big issue for inter-protein contact prediction since it is challenging to find so many interlogs (i.e., interacting homologs). Because of this, currently DCA for inter-protein contact prediction mainly focuses on prokaryotes and mitochondria [1, 2] since it is relatively easy to find interlogs in prokaryotes, but not in eukaryotes with abundant paralogs.

We have developed a deep learning (DL) method for intra-protein contact prediction [3–5], which greatly outperformed DCA and was officially ranked first in CASP12 [6]. Our DL method needs much fewer sequence homologs than DCA to be effective because it makes use of contact occurrence patterns, in addition to co-evolution, for contact prediction. This abstract shows that DL can also work on inter-protein contact prediction, especially for eukaryotes. To avoid overfitting, we do not train our DL model using any protein complex data (i.e., inter-protein contacts), but use our previous DL model trained by only protein chains (i.e., intra-protein contacts) to predict inter-protein contacts.

We propose a new phylogeny-based method to identify interlogs for a putative interacting protein pair, especially for eukaryotes in which some interacting genes may have big genomic distance. Coupled with DL, this new method works better on eukaryotes than genome-based methods employed by Baker [1] and Marks [2].

As shown in Fig. 1, given a pair of putative interacting proteins A and B under prediction, we first build multiple sequence alignments (MSAs) for A and B, respectively. Then we employ genome- and phylogeny-based strategies to concatenate MSA_A and MSA_B into two paired MSAs consisting of only interlogs. Finally, we use our DL method to predict two inter-protein contact maps and average them for final prediction. Our DL method outperforms pure DCA on three large datasets and works on both prokaryotes and eukaryotes. Table 1 shows the performance comparison on Baker's dataset.

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 295–296, 2018.
https://doi.org/10.1007/978-3-319-89929-9

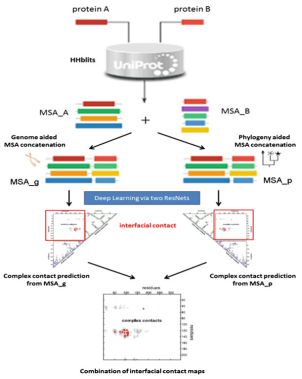

Fig. 1. Method flowchart

Table 1. Inter-protein contact prediction accuracy (%) on Baker's data. GCNN is our method and (s) indicates a web server. EVfold is same as EVcomplex, but run locally with our MSAs. "Genome" and "Phylogeny" denote two MSA generation methods. "Merged" indicates prediction is merged from "Genome" and "Phylogeny". Columns 3–9 show accuracy of top L/10, L/20, 20 and 10 predicted contacts.

Predictor	MSA	L/10	L/20	**20**	**10**
EVcomplex(s)	Built-in	14.25	20.10	21.55	26.55
Gremlin(s)	Built-in	23.74	33.23	41.21	52.76
EVfold	Genome	28.01	39.45	46.90	57.59
EVfold	Phylogeny	15.61	23.09	26.21	36.21
EVfold	Merged	25.13	36.12	42.07	54.83
CCMpred	Genome	28.44	39.54	47.41	53.45
CCMpred	Phylogeny	17.04	25.49	30.34	39.31
CCMpred	Merged	27.70	38.72	46.03	55.52
GCNN	Genome	**51.41**	**60.80**	**62.76**	**68.79**
GCNN	Phylogeny	32.61	39.30	42.24	47.59
GCNN	Merged	48.25	57.09	60.52	65.86

References

1. Ovchinnikov, S., Kamisetty, H., Baker, D.: Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. Elife **3**, e02030 (2014)
2. Hopf, T.A., et al.: Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife **3**, e03430 (2014)
3. Wang, S., et al.: Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput. Biol. **13**(1), e1005324 (2017)
4. Wang, S., et al.: Folding membrane proteins by deep transfer learning. Cell Syst. **5**(3), 202–211. e3 (2017)
5. Wang, S., Sun, S., Xu, J.: Analysis of deep learning methods for blind protein contact prediction in CASP12. Proteins: Struct. Funct. Bioinf. (2017)
6. Schaarschmidt, J., et al.: Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. Proteins (2017)

Author Index

Achtman, Mark 225
Alikhan, Nabil-Fareed 225
Almodaresi, Fatemeh 271
Altenbuchinger, Michael 75
Aluru, Chaitanya 211
Aluru, Srinivas 211

Bakhtiari, Mehrdad 243
Bankevich, Anton 1
Bafna, Vineet 243, 276
Bansal, Vikas 243
Basso, Rebecca Sarto 278
Beltran, Pierre M. Jean 54
Beerenwinkel, Niko 269, 283, 285
Berger, Bonnie 245, 251, 280, 285
Bender, Michael A. 271
Bepler, Tristan 245
Bohmann, Kristine 276
Bonnet, Édouard 248
Boroojeny, Ali Ebrahimpour 37
Brasch, Julia 245

Canzar, Stefan 21
Cannistra, Anthony 263
Chakraborty, Shounak 21
Chikhi, Rayan 105
Cho, Hyunghoon 251
Chitsaz, Hamidreza 37
Chockalingam, Sriram P. 211
Craven, Mark 194
Cristea, Ileana M. 54
Cristea, Simona 283
Crivello, Julianna 293
Crovella, Mark 263

DeCourcy, Alex 138
Durif, G. 254

Eskin, Eleazar 289

Fan, Jason 263
Farrell, Nina 291
Ferdman, Michael 271
Franklin Pugh, B. 291
Fried, Inbar 263

Gagie, Travis 105
Galitzine, Cyril 54
Gallagher, Suzanne Renick 37
Gasch, Audrey 194
Geiger, Tamar 283
Gilbert, David M. 293
Gilbert, M. Thomas P. 276
Görtler, Franziska 75
Gu, Quanquan 293
Gymrek, Melissa 243

Halperin, Eran 274
Hammer, Stefan 256
Hescott, Benjamin 263
Ho, Yi-Hsuan 194
Hochbaum, Dorit S. 278
Hoinka, Jan 287
Hormozdiari, Fereydoun 259
Huynh, Linh 259

Ideker, Trey 266

Jahn, Katharina 269, 285
Johnson, Rob 271
Joseph, Tyler A. 90

Kolmogorov, Mikhail 261
Kuipers, Jack 269, 285
Kuosmanen, Anna 105

Lai, William K. M. 291
Lambert-Lacroix, S. 254
Larson, Gary 122
Leiserson, Mark D. M. 263
Li, Sujun 138
Lim, Tim 263
Lin, Yu 261
Liu, Yang 266
Luhmann, Nina 225
Luo, Yunan 266

Ma, Jian 293
Ma, Jianzhu 266
Mahony, Shaun 291
Mäkinen, Veli 105
Malikic, Salem 269
Marschall, Tobias 21
Mirarab, Siavash 276
Modolo, L. 254
Mold, J. E. 254
Morin, Andrew 245

Noble, Alex J. 245
Numanagić, Ibrahim 280

Oefner, Peter J. 75
O'Neill, Rachel 293
Orenstein, Yaron 154

Pandey, Prashant 271
Patro, Rob 271
Paavilainen, Topi 105
Pe'er, Itsik 90
Peng, Jian 251, 266
Pevzner, Pavel 1, 261
Picard, F. 254
Ponty, Yann 256
Przytycka, Teresa M. 287

Quince, Christopher 225

Rahmani, Elior 274
Roch, Sebastien 167
Rosset, Saharon 274
Rzążewski, Paweł 248

Sankararaman, Sriram 274, 289
Sarmashghi, Shahab 276
Sahinalp, S. Cenk 37, 269
Sasaki, Takayo 293
Schaffner, Thomas 263
Schmidler, Scott 122
Schulz, Marcel H. 21
Schweiger, Regev 274
Shajii, Ariya 280
Shapiro, Lawrence 245
Sharan, Roded 283
Sharifi-Zarchi, Ali 37
Shleizer-Burko, Sharona 243
Shrestha, Akash 37
Sikora, Florian 248
Singer, Jochen 285
Silverbush, Dana 283
Solbrig, Stefan 75
Soulé, Antoine 177
Spang, Rainer 75
Steyaert, Jean-Marc 177
Sverchkov, Yuriy 194
Swiderski, Piotr 287

Tang, Haixu 138
Thankachan, Sharma V. 211
Thorne, Jeffrey L. 122
Tomescu, Alexandru 105

Vandin, Fabio 278
Vitek, Olga 54

Waldispühl, Jérôme 177
Wang, Kun-Chieh 167
Wang, Sheng 295
Wang, Wei 256
Wang, Yijie 287
Wettig, Tilo 75
Will, Sebastian 256
Wu, Yue 289

Xu, Jinbo 295

Yamada, Naomi 291
Yang, Yang 293
Yanovich, Gali 283
Ye, Qing 266
Yuan, Jeffrey 261

Zhou, Tian-Ming 295
Zhou, Zhemin 225