BeWith: A Between-Within Method for Module Discovery in Cancer using Integrated Analysis of Mutual Exclusivity, Co-occurrence and Functional Interactions (Extended Abstract)

Phuong Dao[1], Yoo-Ah Kim[1], Sanna Madan[2], Roded Sharan[3](✉),
and Teresa M. Przytycka[1](✉)

[1] National Center of Biotechnology Information, NLM, NIH, Bethesda, MD, USA
przytyck@ncbi.nlm.nih.gov
[2] Department of Computer Science, University of Maryland, College Park, MD, USA
[3] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel
roded@post.tau.ac.il

The analysis of the cancer mutational landscape has been instrumental in studying the disease and identifying its drivers and subtypes. In particular, mutual exclusivity of mutations in cancer drivers has recently attracted a lot of attention. These relationships can help identify cancer drivers, cancer-driving pathways, and subtypes [1–4]. The co-occurrence of mutations has also provided critical information about possible synergistic effects between gene pairs [5].

Importantly, both properties can arise due to several different reasons, making the interpretation challenging. Specifically, mutually exclusive mutations within a functionally interacting gene module may indicate that a mutation in either of the two genes dysregulates the same pathway. On the other hand, mutually exclusive mutations might reflect a situation where two genes drive different cancer types, which is more likely to occur between genes belonging to different pathways. We have previously observed that within cancer type mutual exclusivity is more enriched with physically interacting genes than between cancer types mutual exclusivity [2]. Thus, the interaction information between genes might provide hints toward the nature of the mutual exclusivity. In addition, mutual exclusivity is not necessarily limited to cancer drivers, and therefore a proper understanding of this property is critical for obtaining a better picture of cancer mutational landscape and for cancer driver prediction.

The co-occurrence of mutations might also emerge due to a number of different causes. Perhaps the most important case is when the co-inactivation of two genes simultaneously might be beneficial for cancer progression such as the co-occurrence of TP53 mutation and MYC amplification [5] or co-occurring mutations in PIK3CA and RAS/KRAS [6]. Alternatively, co-occurrence of somatic mutations might indicate the presence of a common mutagenic process.

Given the diversity of reasons for observing the mutational patterns, we hypothesised that jointly considering co-occurrence, mutual exclusivity and functional interaction relationships will yield a better understanding of the mutational

P. Dao and Y.-A. Kim—Equal contribution.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 370–371, 2017.
DOI: 10.1007/978-3-319-56970-3

landscape of cancer. To address this challenge, we designed a general framework, named BeWith, for identifying modules with different combinations of mutation and interaction patterns. On a high level, BeWith tackles the following problem: given a set of genes and two types of edge scoring functions (within and between scores), find the clusters of genes such that genes within a cluster maximize the within scores while gene pairs in two different clusters maximize the between scores. We formulated the BeWith module identification problem as an Integer Linear Programming (ILP) and solved it to optimality.

In this work, we focused on three different settings of the BeWith framework: BeME-WithFun (mutual exclusivity between different modules and functional similarity of genes within modules), BeME-WithCo (mutual exclusivity between modules and co-occurrence within modules), and BeCo-WithMEFun (co-occurrence between modules while enforcing mutual exclusivity and functional interactions within modules). By utilizing different settings of within and between properties, BeWith revealed complex relations between mutual exclusivity, functional interactions, and co-occurrence. In particular, BeME-WithFun identified functionally coherent modules containing cancer associated genes. By looking for co-occurring mutations inside a module, the BeME-WithCo setting allowed us to investigate mutated modules in a novel way and help uncover synergetic gene pairs in breast cancer. Going beyond cancer driving mutations, the setting also provided insights into underlying mutagenic processes in cancer. Importantly, the BeWith formulation is very general and can be used to interrogate other aspects of the mutational landscape by exploring different combinations of within-between definitions and constraints with simple modifications.

Implementation is available at https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/software/bewith.html.

References

1. Babur, Ö., Gönen, M., Aksoy, B.A., Schultz, N., Ciriello, G., Sander, C., Demir, E.: Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. Genome Biol. **16**, 45 (2015)
2. Kim, Y.A., Cho, D.Y., Dao, P., Przytycka, T.M.: MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. Bioinformatics **31**(12), i284–92 (2015)
3. Leiserson, M.D.M., Blokh, D., Sharan, R., Raphael, B.J.: Simultaneous identification of multiple driver pathways in cancer. PLoS Comput. Biol. **9**(5), e1003054 (2013)
4. Leiserson, M.D.M., Hsin-Ta, W., Fabio, V., Raphael, B.J.: CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. Genome Biol. **16**(1), 72 (2015)
5. Ulz, P., Heitzer, E., Speicher, M.R.: Co-occurrence of MYC amplification and TP53 mutations in human cancer. Nat. Genet. **48**(2), 104–106 (2016)
6. Cancer Genome Atlas Network: Comprehensive molecular portraits of human breast tumours. Nature **490**(7418), 61–70 (2012)

K-mer Set Memory (KSM) Motif Representation Enables Accurate Prediction of the Impact of Regulatory Variants

Yuchun Guo, Kevin Tian, Haoyang Zeng,
and David K. Gifford(✉)

MIT, Computer Science and Artificial Intelligence Laboratory,
Cambridge, MA, USA
gifford@mit.edu

Introduction

The discovery and representation of transcription factor (TF) DNA sequence binding specificities is critical for understanding regulatory networks and interpreting the impact of non-coding genetic variants. The position weight matrix (PWM) model does not represent binding specificities accurately because it assumes that base positions in the motif are independent. Recent studies have shown that DNA sequences proximal to a TF motif core may affect DNA shape and hence TF binding [1]. We hypothesized that a motif model that preserves base positional dependences and includes proximal flanking bases would improve upon existing motif models.

Approach

We introduce the K-mer Set Memory (KSM) model that represents TF binding specificity as a set of aligned gapped and ungapped k-mers that are over-represented at TF binding sites. KSM motif matching requires an exact match of one or more component k-mers, thus preserving inter-positional dependences. The k-mers matching the motif core and the flanking bases are combined non-additively to score the KSM motif matches. We have developed a *de novo* motif discovery method called KMAC to learn KSM and corresponding PWM motifs from TF ChIP-seq data.

Results

We compared KMAC with four state-of-the-art motif discovery methods, MEME, MEME-chip, Homer, and Weeder2, on discovering PWM motifs from the binding sites of 78 TFs in 209 ENCODE ChIP-seq experiments. KMAC identified previously published PWM motifs in more experiments than the other methods.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 372–374, 2017.
DOI: 10.1007/978-3-319-56970-3

We then compared the performance of KSM versus other motif models in predicting *in vivo* TF binding by discriminating TF-bound sequences from unbound sequences. For the GAPB dataset, a KSM motif outperforms PWM motifs computed by KMAC, MEME, and Homer, as well as representations that model base inter-dependences such as TFFM [2] and Slim [3] (Fig. 1A). For sequences with identical PWM scores, the KSM scores of the positive sequences are generally higher than those of the negative sequences (Fig. 1B). This is because KSM motif matches in positive sequences often contain more KSM k-mers that cover motif flanking bases than those in negative sequences.

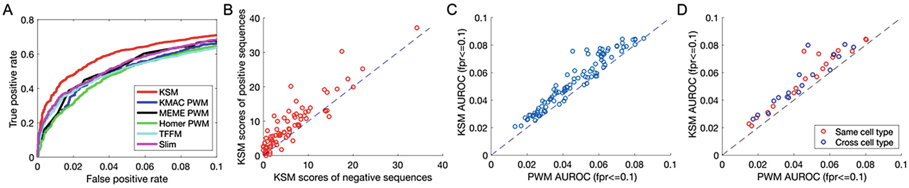

Fig. 1. KSM outperforms PWM in predicting *in vivo* TF binding in held-out data.

Out of 104 ChIP-seq datasets, KSMs outperform PWMs in 85 datasets; and PWMs do not outperform KSMs in any dataset (Fig. 1C). Overall, a KSM significantly outperforms a PWM ($p = 4.79e\text{-}18$, paired Wilcoxon signed rank test) and a TFFM in predicting TF binding ($p = 0.045$, paired Wilcoxon signed rank test). We also found that a KSM is able to generalize across cell types. For 19 unique TFs, KSMs significantly outperformed PWMs when a motif learned from one cell type (K562) is used to predict binding of the same TF in another cell type (GM12878 or H1-hESC) (Fig. 1D). The KSM predictions across the cell types perform similarly to the KSM predictions in the same cell type ($p = 0.091$, paired Wilcoxon signed rank test).

(A) The partial ROC (fpr <= 0.1) of KSM and other models for predicting ChIP-seq binding of GABP in K562 cells. (B) Comparison of the mean KSM scores of positive versus negative sequences that corresponds to the same PWM scores. Each point represents a set of sequences that have the same PWM score. (C) Comparison of the median AUROC (fpr <= 0.1) scores of KSM and PWM for 104 experiments with five cross-validation datasets. (D) Similar to (C), but comparing KSM and PWM in the same cell type (red) or cross cell type (blue) in 19 TFs.

In addition, evaluated the ability of different sequence features to predict the regulatory activity of e-QTLs using a computational framework [4] that performed the best in the CAGI 4 “eQTL-causal SNPs” challenge. We found that KSM derived features (AUPRC = 0.461, AUROC = 0.683) outperformed Homer PWM derived features (AUPRC = 0.434, AUROC = 0.629), MEME PWM derived features (AUPRC = 0.408, AUROC = 0.619) and sequence features derived from DeepSEA [5], a deep learning model (AUPRC = 0.396, AUROC = 0.628), in predicting the differential regulatory activities of e-QTL alleles. The combined KSM and DeepSEA features achieved the best performance (AUPRC = 0.464, AUROC = 0.696).

Finally, we have created a public resource of KSM and PWM motifs from more than one thousand ENCODE TF ChIP-seq datasets.

Conclusion

We found that the K-mer Set Model (KSM) is a more powerful motif representation than the PWM and TFFM models for identifying held-out DNA sequences that are bound by a TF. We also found that KMAC more accurately discovers PWM motifs than other tested methods. Thus, the KSM and PWM models produced by the KMAC method improve the ability to model TF binding specificities, and enable more accurate characterization of non-coding genetic variants.

References

1. Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., Mann, R.S.: Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell **147**, 1270–1282 (2011)
2. Mathelier, A., Wasserman, W.W.: The next generation of transcription factor binding site prediction. PLoS Comput. Biol. **9**, e1003214 (2013)
3. Keilwagen, J., Grau, J.: Varying levels of complexity in transcription factor binding motifs. Nucl. Acids Res. **43**, e119–e119 (2015)
4. Zeng, H., Edwards, M.D., Guo, Y., Gifford, D.K.: Accurate eQTL prioritization with an ensemble-based framework. bioRxiv. 69757 (2016)
5. Zhou, J., Troyanskaya, O.G.: Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Meth. **12**, 931–934 (2015)

Network-Based Coverage of Mutational Profiles Reveals Cancer Genes

Borislav H. Hristov and Mona Singh(✉)

Department of Computer Science and Lewis-Sigler Institute for Integrative Genomics,
Princeton University, Princeton, NJ, USA
mona@cs.princeton.edu

Summary. A central goal in cancer genomics is to identify the somatic alterations that underpin tumor initiation and progression. This task is challenging as the mutational profiles of cancer genomes exhibit vast heterogeneity, with many alterations observed within each individual, few shared somatically mutated genes across individuals, and important roles in cancer for both frequently and infrequently mutated genes. While commonly mutated cancer genes are readily identifiable, those that are rarely mutated across samples are difficult to distinguish from the large numbers of other infrequently mutated genes. Here, we introduce a method that considers per-individual mutational profiles within the context of protein-protein interaction networks in order to identify small connected subnetworks of genes that, while not individually frequently mutated, comprise pathways that are perturbed across (i.e., "cover") a large fraction of the individuals. We devise a simple yet intuitive objective function that balances identifying a small subset of genes with covering a large fraction of individuals. We show how to solve this problem optimally using integer linear programming and also give a fast heuristic algorithm that works well in practice. We perform a large-scale evaluation of our resulting method, nCOP, on 6,038 TCGA tumor samples across 24 different cancer types. We demonstrate that our approach is more effective in identifying cancer genes than both methods that do not utilize any network information as well as state-of-the-art network-based methods that aggregate mutational information across individuals. Overall, our work demonstrates the power of combining per-individual mutational information with interaction networks in order to uncover genes functionally relevant in cancers, and in particular those genes that are less frequently mutated.

Methods. We model the biological network as an undirected graph G where each vertex represents a gene, and there is an edge between two vertices if an interaction has been found between the corresponding proteins. We annotate each node in the network with the IDs of the individuals having one or more mutations in the corresponding gene. Our goal is to find a relatively small connected component G' such that most patients have mutations in one of the genes within it. A small subgraph is more likely to consist of functionally related genes and is less likely to be the result of overfitting to the set of individuals whose diseases we are analyzing. However, we would also like our model to have the greatest possible explanatory power—that is, to account for, or cover, as many patients as possible by including genes that are mutated within their cancers.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 375–376, 2017.
DOI: 10.1007/978-3-319-56970-3

We formulate our problem to balance these two competing objectives with a parameter α that controls the trade-off between keeping the subgraph small and covering more patients as to minimize $\alpha X + (1 - \alpha) Size(G')$, where X is the fraction of patients that do not have an alteration in a gene included in G' (i.e., they are uncovered) and $Size(G')$ is the size of the subgraph.

For a fixed value of α, we have developed two approaches to solve the underlying optimization problem: one based on linear programming and the other a fast greedy heuristic. To select an appropriate α for a set of cancer samples, we devise a simple but effective data-driven cross-validation technique. In particular, we split our samples into training, validation and test sets. A test set of (10%) of the patients is completely withheld. While varying α in small increments in the interval $(0; 1)$, the remaining data is repeatedly split (100 times for each value of α) into training (80%) and validation (20%) sets. For each split, the algorithm is run on the training set to find G'. The fractions of patients covered (by the selected G') in the training and validation sets are compared. The parameter α is selected where performance on the validation sets deviates as compared to the training sets. Once α is chosen for a set of cancer samples, we repeatedly (1000 times) run the algorithm on this set, each time withholding a fraction (15%) of the patients in order to introduce some randomness in the process. Genes are then ranked by the number of times they appear in G'.

Results. We run nCOP on somatic point mutation data from 24 different TCGA cancer types. We show that nCOP effectively uses network information to uncover known cancer genes by considering how well it recapitulates known cancer genes (CGCs) in comparison to network-agnostic methods. We find that nCOP outperforms MutSigCV 2.0 (Lawrence et al. 2013), a state-of-the-art frequency-based approach, on 21 of the 24 cancer types, and a basic set cover approach on all 24 types. We also show that nCOP is more effective in uncovering known cancer genes than Muffinn (Cho et al. 2016), a recent network-based method that considers mutations found in interacting genes. Finally, we examine the non-CGC genes which are highly ranked by nCOP and observe that they tend to be less frequently mutated. Our results are consistent across the three different networks we used (HPRD, HINT, and Biogrid), showing the robustness of the method with respect to the underlying network. Further, we demonstrate that our training-validation-test set framework is a highly effective approach for choosing an α that balances patient coverage with subnetwork size.

In summary, we present nCOP, a method that incorporates individual mutational profiles with protein–protein interaction networks, and show it is a powerful approach for uncovering cancer genes. Researchers can use our framework to rapidly and easily prioritize cancer genes, as nCOP requires only straightforward inputs and runs on a desktop machine. Indeed, nCOP's efficiency, robustness, and ease of use make it an excellent choice to investigate cancer as well as possibly other complex diseases. nCOP can be freely downloaded at: http://compbio.cs.princeton.edu/ncop/.

Ultra-Accurate Complex Disorder Prediction: Case Study of Neurodevelopmental Disorders

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

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

Motivation and Problem Definition. Early prediction of complex disorders (e.g. autism, intellectual disability or schizophrenia) is one of the main goals of personalized genomics and precision medicine. Considering the high genetic heritability of neurodevelopmental disorders ($h^2 > 0.5$ for autism [1]) we are proposing a novel problem and framework for accurate prediction of autism and related disorders based on rare and *de novo* genetic variants [2]. However, a positive diagnosis/prediction of a complex disorder (e.g., autism or intellectual disability) can have a severe negative psychological and economical impact on affected individuals and their family. Thus, one of the primary practical constraints in developing models and methods for prediction of a severe complex disorder *is to guarantee a false positive prediction/discovery rate (FDR) of virtually zero*. Hence, we are introducing a novel problem for prediction of complex disorders for a subset of affected cases with very low false positive prediction. We denote this problem as Ultra-Accurate Disorder Prediction (UADP) problem.

Methods. We have proposed framework for solving the UADP problem denoted as Odin (**O**racle for **DI**sorder predictio**N**). Odin will intuitively predict an input/test sample to be an affected case if and only if it satisfies two conditions:

1. The input sample is "far" from any unaffected control sample
2. The input sample is "close" to many affected case samples

For satisfying the first condition, we simply use the nearest neighbor (NN) approach using a distance function (e.g., Euclidean distance). For satisfying the second condition, we first develop a novel algorithm that finds a cluster (together with a dimension reduction) that contains a significant number of affected cases and does not contain any unaffected controls. This cluster is denoted as *unicolor cluster*, as it only includes the affected cases. We denote the problem of finding such a cluster as Unicolor Clustering with Dimensionality Reduction (UCDR) problem. An input sample passes the second condition if it falls inside of this unicolor cluster. A weighted version of UCDR, where we can assign weights to each dimension is denoted as Weighted Unicolor Clustering with Dimensionality Reduction (WUCDR) problem. We have shown that the decision version of an UCDR instance is NP-complete using reduction from equal-subset sum problem [3]. We propose an iterative approach with two steps to solve the WUCDR problem. In the first step, given weights for each dimension, we find the cluster to

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 377–378, 2017.
DOI: 10.1007/978-3-319-56970-3

cover a maximum number of affected cases. In the second step, given the cluster from the first step, we find the new set of weights for each dimension by using a linear programming (LP) formulation.

Results. We used the leave-one-out (LOO) cross validation technique to compare the prediction power of Odin and the of k-NN and SVM classifiers. As our stated goal is to keep the false positive prediction of unaffected samples as cases close to zero, we will only consider the most conservative results for each method (i.e., where false discovery rate (FDR) < 0.01). For the same FDR threshold, Odin's true positive rate for predicting autism is at least twice higher than the best k-NN result (for various values of k) and significantly higher than SVM. Our experimental results indicate the ability of our approach in ultra-accurate prediction of autism spectrum disorder (ASD) in additional 8% of cases which do not have a severe mutation in *recurrently* mutated genes, with less than 0.5% of false positive prediction rate for unaffected controls.

Odin is publicly available at https://github.com/HormozdiariLab/Odin.

References

1. Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H., Hultman, C.M., Reichenberg, A.: The familial risk of autism. JAMA **311**(17), 1770–1777 (2014)
2. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al.: The contribution of de novo coding mutations to autism spectrum disorder. Nature **515**(7526), 216–221 (2014)
3. Woeginger, G.J., Yu, Z.: On the equal-subset-sum problem. Inf. Process. Lett. **42**(6), 299–302 (1992)

Inference of the Human Polyadenylation Code

Michael K.K. Leung[1,2(✉)], Andrew Delong[1,2],
and Brendan J. Frey[1,2,3]

[1] Deep Genomics, MaRS Centre, Heritage Building, Suite 320,
Toronto, ON M5G 1L7, Canada
mleung@psi.toronto.edu
[2] Department of Electrical and Computer Engineering, University of Toronto,
Toronto M5S 3G4, Canada
[3] Banting and Best Department of Medical Research, University of Toronto,
Toronto M5S 3E1, Canada

Abstract. Processing of transcripts at the 3'-end is a two-step procedure that involves cleavage at a polyadenylation site followed by the addition of a poly (A)-tail. By selecting which polyadenylation site is cleaved in transcripts with multiple sites, alternative polyadenylation enables genes to produce transcript isoforms with different 3'-ends. To facilitate the identification and treatment of disease-causing mutations that affect polyadenylation and to understand the underlying regulatory processes, a computational model that can accurately predict polyadenylation patterns based on genomic features is desirable. Previous works have focused on identifying candidate polyadenylation sites, as well as classifying sites which may be tissue-specific. However, what is lacking is a predictive model of the underlying mechanism of site selection, competition, and processing efficiency in a tissue-specific manner. We develop a deep learning model that trains on 3'-end sequencing data and predicts tissue-specific site selection among competing polyadenylation sites in the 3' untranslated region of the human genome.

Two neural network architectures are evaluated: one built on hand-engineered features, and another that directly learns from the genomic sequence. The hand-engineered features include polyadenylation signals, cis-regulatory elements, n-mer counts, nucleosome occupancy, and RNA-binding protein motifs. The direct-from-sequence model is inferred without prior knowledge on polyadenylation, based on a convolutional neural network trained with genomic sequences surrounding each polyadenylation site as input. Both models are trained using the Tensor Flow library.

The proposed polyadenylation code can predict functional site selection among competing polyadenylation sites across all tissues. Importantly, it does so without relying on evolutionary conservation. The model can directly distinguish pathogenic from benign variants that appear near annotated polyadenylation sites, achieving a classification AUC of 0.98 ($p < 1 \times 10^{-8}$) on ClinVar. We further demonstrate the potential use of the same model to predict the effects of antisense oligonucleotides to redirect polyadenylation and to scan the genome to find candidate polyadenylation sites.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, p. 379, 2017.
DOI: 10.1007/978-3-319-56970-3

Folding Membrane Proteins by Deep Transfer Learning

Zhen Li[1,3], Sheng Wang[1,2], Yizhou Yu[3], and Jinbo Xu[1](✉)

[1] Toyota Technological Institute at Chicago, Chicago, USA
jinboxu@gmail.com
[2] Department of Human Genetics, University of Chicago, Chicago, USA
[3] Department of Computer Science, University of Hong Kong,
Hong Kong, China

Membrane proteins (MPs) are important for drug design and have been targeted by approximately half of current therapeutic drugs. In many genomes 20–40% of genes encode MPs. In particular, Human genome has >5,000 reviewed MPs and more than 3000 of them are non-redundant at 25% sequence identity. Experimental determination of MP structures is challenging as they are often too large for NMR experiments and very difficult to crystallize. As of October 2016, there are ~510 non-redundant MPs with solved structures, and a majority number of MPs have no solved structures.

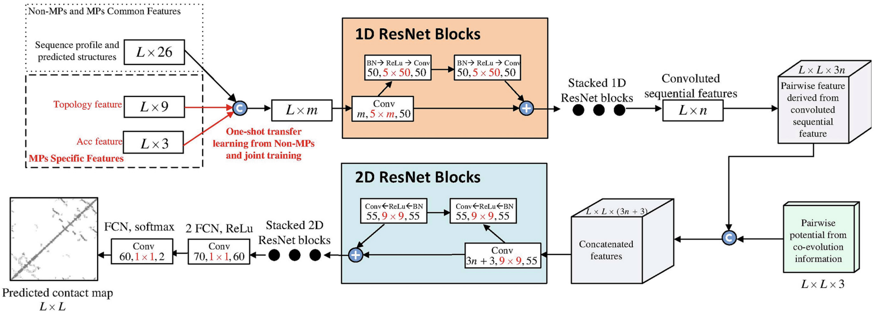

Fig. 1. Overview of our deep learning model for MP contact prediction where L is the sequence length of one MP under prediction.

Developing computational methods for MP structure prediction is challenging partially due to lack of MPs with solved structures for homology modeling or for parameter estimation of ab initio folding. Recently contact-assisted ab initio folding has made good progress [1]. This technique first predicts the contacts of a protein and then use predicted contacts as restraints to guide folding simulation. Contact-assisted folding heavily depends on accurate prediction of protein contacts. Co-evolution analysis can predict contacts accurately for some proteins with a large number of sequence homologs. However, protein families without good templates in PDB on average have many fewer sequences homologs than those with good templates.

Zhen Li and Sheng Wang – These authors contributed equally to the work as first authors.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 380–382, 2017.
DOI: 10.1007/978-3-319-56970-3

We have developed a deep transfer learning method that can significantly improve MP contact prediction by learning contact occurrence patterns from thousands of non-membrane proteins (non-MPs). We treat a contact map as an image and formulate contact prediction similarly as pixel-level image labeling. As shown in Fig. 1, our deep network is composed of two concatenated deep residual neural networks. Each network consists of some residual blocks and each block has 2 convolution and ReLU layers. The first residual network conducts 1-dimensional (1D) convolutional transformations of sequential features. Its output is converted to a 2-dimensional (2D) matrix by an operation called outer concatenation and fed into the 2nd residual network together with the pairwise features. The 2nd residual network conducts 2D convolutional transformations of its input and feeds its output into logistic regression, which predicts the probability of any two residues in a contact. We predict all the contacts of a protein simultaneously to capture contact occurrence patterns and improve prediction accuracy. We use two types of protein features: sequential features and pairwise features. The sequential features include protein sequence profile, secondary structure and solvent accessibility predicted by RaptorX-Property [2]. The pairwise features include co-evolutionary strength generated by CCMpred [3], mutual information and pairwise contact potential. Some MP-specific features are also tested.

We studied three training strategies: NonMP (only non-MPs used as training proteins), MP-only (only MPs used as training proteins) and Mixed (both non-MPs and MPs used as training proteins). Tested on 510 non-redundant MPs, our deep models trained by NonMP only, MP-only and Mixed have top L/10 long-range prediction accuracy 0.69, 0.63 and 0.72, respectively, all much better than CCMpred (0.47) and the CASP11 winner MetaPSICOV (0.55). When only contacts in transmembrane regions are evaluated, our models have top L/10 long-range accuracy 0.57, 0.53, and 0.62, respectively, again much better than MetaPSICOV (0.45) and CCMpred (0.40). These results suggest that sequence-structure relationship learned by our deep model from non-MPs generalizes well to MP contact prediction and that non-MPs and MPs share common contact occurrence patterns.

Improved contact prediction also leads to better contact-assisted folding. We build 3D structure models for a MP by feeding its top predicted contacts to the CNS package, and evaluate model quality by TMscore, which ranges from 0 to 1, indicating the worst and best models, respectively. A model with TMscore >0.5 (0.6) is (very) likely to have a correct fold. The average TMscore (RMSD in Å) of the 3D models built by our three models MP-only, NonMP-only and Mixed are 0.45 (14.9), 0.49 (13.2), and 0.52 (10.8), respectively. By contrast, the average TMscore (RMSD in Å) of the 3D models built from MetaPSICOV and CCMpred-predicted contacts are 0.39 (16.7) and 0.36 (17.0), respectively. When the best of top 5 models are considered and TMscore = 0.6 is used as cutoff, our three models can predict correct folds for 110, 160, and 200 of 510 MPs, respectively, while MetaPSICOV and CCMpred can do so for only 77 and 56 of them, respectively. Homology modeling can correctly fold 41 MPs when MPs and non-MPs are used as templates and 3 MPs when only non-MPs are used as templates. When TMscore = 0.5 is cutoff, our Mixed method, MetaPSICOV, and CCMpred can predict correct folds for 283, 147, and 122 MPs, respectively.

References

1. Wang, S., Sun, S., Li, Z., Zhang, R., Xu, J.: Accurate De Novo prediction of protein contact map by ultra-deep learning model. PLoS Comput. Biol. **13**, e1005324 (2017)
2. Wang, S., Li, W., Liu, S., Xu, J.: RaptorX-Property: a web server for protein structure property prediction. Nucleic Acids Res. gkw306 (2016)
3. Seemayer, S., Gruber, M., Söding, J.: CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics **30**, 3128–3130 (2014)

A Network Integration Approach for Drug-Target Interaction Prediction and Computational Drug Repositioning from Heterogeneous Information

Yunan Luo[1,3], Xinbin Zhao[2], Jingtian Zhou[2], Jinling Yang[1], Yanqing Zhang[1],
Wenhua Kuang[2], Jian Peng[3(✉)], Ligong Chen[2(✉)], and Jianyang Zeng[1(✉)]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University,
Beijing, China
zengjy321@tsinghua.edu.cn
[2] School of Pharmaceutical Sciences, Tsinghua University, Beijing, China
ligongchen@biomed.tsinghua.edu.cn
[3] Department of Computer Science, University of Illinois at Urbana-Champaign,
Champaign, USA
jianpeng@illinois.edu

The emergence of large-scale genomic, chemical and pharmacological data provides new opportunities for drug discovery and repositioning. Systematic integration of these heterogeneous data not only serves as a promising tool for identifying new drug-target interactions (DTIs), which is an important step in drug development, but also provides a more complete understanding of the molecular mechanisms of drug action. In this work, we integrate diverse drug-related information, including drugs, proteins, diseases and side-effects, together with their interactions, associations or similarities, to construct a heterogeneous network with 12,015 nodes and 1,895,445 edges. We then develop a new computational pipeline, called DTINet, to predict novel drug-target interactions from the constructed heterogeneous network. Specifically, DTINet focuses on learning a low-dimensional vector representation of features for each node, which accurately explains the topological properties of individual nodes in the heterogeneous network, and then predicts the likelihood of a new DTI based on these representations via a vector space projection scheme. DTINet achieves substantial performance improvement over other state-of-the-art methods for DTI prediction. Moreover, we have experimentally validated the novel interactions between three drugs and the cyclooxygenase (COX) protein family predicted by DTINet, and demonstrated the new potential applications of these identified COX inhibitors in preventing inflammatory diseases. These results indicate that DTINet can provide a practically useful tool for integrating heterogeneous information to predict new drug-target interactions and repurpose existing drugs.

The full paper of DTINet is available at [1]. The source code of DTINet and the input heterogeneous network data can be downloaded from https://github.com/luoyunan/DTINet.

Y. Luo et al.—These authors contributed equally to this work.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 383–384, 2017.
DOI: 10.1007/978-3-319-56970-3

Acknowledgments. This work was supported in part by the National Basic Research Program of China (Grant 2011CBA00300 and 2011CBA00301), the National Natural Science Foundation of China (Grant 61033001, 61361136003, 61472205 and 81470839), the China's Youth 1000-Talent Program, the Beijing Advanced Innovation Center for Structural Biology, and the Tsinghua University Initiative Scientific Research Program (Grant. 20161080086). J.P. received support as an Alfred P. Sloan Research Fellow. We acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

Reference

1. Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., Zeng, J.: A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information (2017). bioRxiv. doi:https://doi.org/10.1101/100305

Epistasis in Genomic and Survival Data of Cancer Patients

Dariusz Matlak and Ewa Szczurek(✉)

Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw, Warsaw, Poland
szczurek@mimuw.edu.pl

Extended Abstract

Fitness is a measure of replicative and survival success of an individual, relative to competitors in the same population. Epistasis is an interaction between genes, and refers to departure from independence of effects that their genomic alterations have on fitness. Beerenwinkel et al. [2] defined epistatic interactions not only among two, but also more genes. Here, we consider epistasis of genes in their contribution to fitness of tumors in cancer patients.

Current state of the art cancer therapies have limited efficacy due to toxicity and rapid development of drug resistance. Recently, therapies exploiting synthetic lethal interactions between genes were proposed to overcome these difficulties [8]. Synthetic lethality occurrs when the co-inactivation of two genes results in cellular death, while inactivation of each individual gene is viable. In cancer, one gene inactivation can already occur via the endogenous mutation in the tumor cells, and not in the normal cells of the body. Thus, applying a drug that targets the synthetic lethal partner of that gene will selectively kill cancer cells, leaving the rest viable. A famous example is the interaction between *BRCA1* and *PARP1*. In *BRCA1* deficient cells, treatment with a PARP inhibitor, such as Olaparib [4], is expected to result in selective tumor cell death.

Synthetic lethality is, however, context dependent. For example, compared to the dramatic effect that PARP1 inhibition has on *BRCA1*-deficient cell lines, the efficacy of Olaparib therapy on patients was low, since a positive response was observed in less than 50% of BRCA-mutated cancers [3]. This raises the crucial issue of therapeutic biomarkers. For *BRCA1* and *PARP1*, mutation of *TP53BP1* in addition to *BRCA1* was observed to alleviate the synthetic lethal effect [1]. Thus, in *TP53BP1* deficient tumors, administrating Olaparib is not justified, and unaltered *TP53BP1* is a biomarker of this therapy. Such dependence of pairwise interaction on the mutational status of a third gene is represented by conditional epistasis, a type of the triple epistatic interactions [2].

Experimental approaches to identification of synthetic lethality in human cancer are overwhelmed by the number of, and thus test only small subsets of all possible interactions [7]. The effort and money required for these experiments calls for a pre-selection of synthetic lethal partners based on the computational analysis of existing data. Previous methods [5, 9] aimed at deciphering synthetic lethality from somatic alteration, expression or survival data of cancer patients.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 385–386, 2017.
DOI: 10.1007/978-3-319-56970-3

Here, we introduce SurvLRT, an approach for identification of epistatic gene pairs and triplets in human cancer. We propose a statistical model based on Lehman alternatives [6], which allows to estimate fitness of tumors with a given genotype from survival of carrier patients. We assume that a decrease of fitness of tumors due to a particular genotype is exhibited by a proportional increase of survival of the patients. Based on these assumptions, we introduce a likelihood ratio test for the significance of a given pairwise or triple epistatic interaction. In the test, the null model assumes that there is no epistasis and the gene alterations are independent, while the alternative assumes otherwise. The approach can detect both positive and negative interactions. Compared to our previous approach [9], SurvLRT offers a more natural interpretation of the notion of fitness, as well as a direct statistical test for the significance of epistasis. We analyze the sensitivity and power of SurvLRT in a controlled setting of simulated data. Next, we show that, compared to previous methods, our method performs favorably in predicting known pairwise synthetic lethal interactions. Finally, we apply SurvLRT to detect therapeutic biomarkers, first by recapitulating *TP53BP1*, the known biomarker for therapies based on the *BRCA1*, *PARP1* interaction, and second by identifying a genomic region deleted in tumors as a new and even more significant biomarker.

Acknowledgement. This work was partially supported by the Polish National Science Center grant numbers 2015/16/W/NZ2/00314 and 2015/19/P/NZ2/03780.

References

1. Aly, A., Ganesan, S.: BRCA1, PARP, and 53BP1: conditional synthetic lethality and synthetic viability. J. Mol. Cell Biol. **3**(1), 66–74 (2011)
2. Beerenwinkel, N., Pachter, L., Sturmfels, B.: Epistasis and shapes of fitness landscapes. Stat. Sinica **17**, 1317–1342 (2007)
3. Chan, S.L., Mok, T.: PARP inhibition in BRCA-mutated breast and ovarian cancers. Lancet **376**(9737), 211–213 (2010)
4. Hutchinson, L.: Targeted therapies: PARP inhibitor olaparib is safe and effective in patients with BRCA1 and BRCA2 mutations. Nat. Rev. Clin. Oncol. **7**(10), 549 (2010)
5. Jerby-Arnon, L., Pfetzer, N., Waldman, Y.Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P.A., Gottlieb, E., Ruppin, E.: Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell **158**(5), 1199–1209 (2014)
6. Lehman, E.L.: The power of rank tests. Ann. Math. Statist. **24**(1), 23–43 (1953)
7. Lord, C.J., McDonald, S., Swift, S., Turner, N.C., Ashworth, A.: A high-throughput RNA interference screen for DNA repair determinants of PARP inhibitor sensitivity. DNA Repair (Amst.) **7**(12), 2010–2019 (2008)
8. Porcelli, L., Quatrale, A.E., Mantuano, P., Silvestris, N., Brunetti, A.E., Calvert, H., Paradiso, A., Azzariti, A.: Synthetic lethality to overcome cancer drug resistance. Curr. Med. Chem. **19**(23), 3858–3873 (2012)
9. Szczurek, E., Misra, N., Vingron, M.: Synthetic sickness or lethality points at candidate combination therapy targets in glioblastoma. Int. J. Cancer **133**(9), 2123–2132 (2013)

Ultra-Fast Identity by Descent Detection in Biobank-Scale Cohorts Using Positional Burrows-Wheeler Transform

Ardalan Naseri[1], Xiaoming Liu[2], Shaojie Zhang[1](✉), and Degui Zhi[3]

[1] Department of Computer Science, University of Central Florida,
Orlando, FL 32816, USA
shzhang@cs.ucf.edu

[2] Department of Epidemiology, Human Genetics and Environmental Science,
University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[3] School of Biomedical Informatics,
University of Texas Health Science Center at Houston, Houston, TX 77030, USA
degui.zhi@uth.tmc.edu

Recent advancements in genome-wide SNP array and whole genome sequencing technologies have led to the generation of enormous amounts of population genotype data. Understanding the genetic relationships based on individuals' genotypes will shed light on better insight into precision medicine or population genetics. A basic measure of genetic relationship is Identity by Descent (IBD). IBD is defined as chromosomal segments shared between two individual chromosomes which have been inherited from a common ancestor.

Most of the existing methods for IBD detection, such as IBDseq [1], PLINK [4], and PARENTE [5], can handle both genotype and haplotype data, but they rely on pairwise comparison of all individuals and therefore are not scalable for large number of individuals. GERMLINE [3] avoids pairwise comparison by using hash table on haplotype sequences. Under the assumption that the number of seed matches between any individual and others is constant, the complexity of GERMLINE will grow linearly with the number of samples. However, the individuals in a population share different levels of common substructures, therefore the seed matches may deviate from its idealized linear behavior in a sample with a large number of individuals. As a result, it will not be fast enough for hundreds of thousands of individuals and millions of variant sites.

In this work, we present an efficient computational method, named RaPID (Random Projection for IBD Detection), to find IBD segments larger than a given length in haplotype data. We use an efficient population genotype index, Positional Burrows-Wheeler Transform (PBWT) [2], that scales up linearly with the sample size. PBWT algorithm can compute all haplotype matches that exceed a given length in $O(max(MN, I))$, where M denotes the number of individuals, N the number of variant sites and I the number of matches. The key idea behind PBWT is to sort the sequences by their reversed prefix at each position. The PBWT algorithm sweeps through the list of variant sites and keeps the starting position of each match between neighboring prefixes. PBWT searches for exact matches and cannot tolerate mismatches that might be due to genotyping

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 387–388, 2017.
DOI: 10.1007/978-3-319-56970-3

or phasing errors. In order to account for genotyping or phasing errors, we build PBWT over random projections of the original sequences. We divide the panel into non-overlapping windows with the same length. The length is defined in terms of consecutive variant sites. For each window, we select a variant site at random and find all exact matches that exceed a minimum length using PBWT. We repeat the random projection PBWT multiple times to increase the detection power. Since the error rate is presumably low, a true IBD segment will have a high probability to be identified in some of the multiple runs. On the other hand, a non-IBD segment will have a lower probability to be selected in multiple runs. We model these probabilities as binomial distributions. A matched segment between any two individuals is considered to be an IBD if it was selected more than a certain number of times among a total number of PBWT runs.

To evaluate the performance of RaPID, we have computed the accuracy and power in a simulated population and compared the results with GERMLINE and IBDseq. Accuracy is defined as the percentage of the correctly detected IBD segments which overlap at least 50% with a true IBD. Power is defined as the average of detected proportions of true IBDs. RaPID maintains comparable accuracy and power to GERMLINE and IBDseq while being orders of magnitudes faster than GERMLINE. On our simulated data, the running time of RaPID was more than 100 times faster than GERMLINE when searching for IBDs with a minimum length of 3 cM. Therefore, RaPID would be an appropriate tool for IBD detection in very large Biobank-scale genotyped cohorts. To demonstrate the utility of RaPID for real data, we have applied it on the 1000 Genome Project data. The results show that RaPID can detect population events at different time scales. Implementation is available at https://github.com/ZhiGroup/RaPID.

References

1. Browning, B.L., Browning, S.R.: Detecting identity by descent and estimating genotype error rates in sequence data. Am. J. Hum. Genet. **93**(5), 840–851 (2013)
2. Durbin, R.: Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics **30**(9), 1266–1272 (2014)
3. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., Pe'er, I.: Whole population, genome-wide mapping of hidden relatedness. Genome Res. **19**(2), 318–326 (2009)
4. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., Sham, P.C.: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. **81**(3), 559–575 (2007)
5. Rodriguez, J.M., Bercovici, S., Huang, L., Frostig, R., Batzoglou, S.: Parente2: a fast and accurate method for detecting identity by descent. Genome Res. **25**(2), 280–289 (2015)

Joker de Bruijn: Sequence Libraries to Cover All *k*-mers Using Joker Characters

Yaron Orenstein[1], Ryan Kim[2], Polly Fordyce[3], and Bonnie Berger[1](✉)

[1] Massachusetts Institute of Technology, Cambridge, MA 02139, USA
bab@mit.edu
[2] Research Science Institute, McLean, VA 22207, USA
[3] Stanford University, Stanford, CA 94305, USA

1 Introduction

Protein-DNA, -RNA and -peptide interactions drive nearly all cellular processes. Due to their high importance, high-throughput technologies using sequence libraries that cover all k-mers (i.e. words of length k) have been developed to measure them in a universal and unbiased manner [1]. These techniques all face a similar challenge: the space on the experimental device is limited, restricting the total sequence space that can be probed in a single experiment. While de Bruijn sequences cover all k-mers in the most compact manner, they remain $|\Sigma|^k$ characters long (where Σ is the alphabet, e.g. {A,C,G,T}). Here, we introduce a novel idea and algorithm for sequence design to cover all possible k-mers with a significantly smaller experimental sequence library by using *joker* characters, which represent all characters in the alphabet. Experimentally, such joker characters can be easily incorporated during oligonucleotide or peptide synthesis by using degenerate mixtures of nucleotides or amino acids, at no extra cost. However, joker characters introduce degeneracy which could potentially lower the statistical robustness of the measurements (as a measurement of a single oligonucleotide is now assigned to multiple sequences instead of just one). To address this challenge, we limit the use of joker characters to either one or two joker characters per k-mer, enabling the coverage of $(k+2)$-mers at the same cost and space of k-mers — a savings of a factor of $|\Sigma|^2$ in sequence length (16 and 400 for DNA and amino acid alphabets, respectively). We validate that the library remains capable of *de novo* identification of high-affinity k-mers by testing it on known DNA-protein binding data for hundreds of proteins. The implementation of our algorithm is freely available at `jokercake.csail.mit.edu`.

2 Methods

We propose a novel solution to the problem of generating a short sequence covering all k-mers using joker characters. The solution is based on two steps: (i) a greedy heuristic; (ii) and an ILP formulation. The greedy heuristic examines at each step an addition of $k-1$ characters from Σ followed by a joker character. The addition that covers the most k-mers that are yet to be covered p times

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 389–390, 2017.
DOI: 10.1007/978-3-319-56970-3

is chosen and added to the current sequence. The algorithm terminates when all k-mers have been covered at least p times. The ILP formulation minimizes the number of k-mers in the sequence under two sets of constraints. The first requires that all k-mers occur at least p times. The second guarantees that the k-mer occurrences can form a sequence. The ILP is solved using Gurobi ILP solver version 6.5.2 [2], where it is given the greedy solution as a starting solution.

3 Results

To test the performance of our algorithm, we ran it on different parameter combinations. We ran the greedy heuristic on $5 \leq k \leq 8$ for a DNA alphabet and $3 \leq k \leq 4$ for an amino acid alphabet, with $p = 1$. We then ran the ILP solver, starting from the greedy solution, with a time limit of 4 weeks. Results show that the greedy algorithm produces a sequence that is much smaller than the original de Bruijn sequence; i.e., less than 40% and 8% of the original for DNA and amino acid alphabets, respectively. Following the ILP solver, sequence length drops even further to less than 33% and 8% of the original, respectively, where the theoretical lower bounds are 25% and 5%, respectively. To test the performance of our algorithm in covering k-mers multiple times, we ran the greedy heuristic on $k = 6$, a DNA alphabet, and $1 \leq p \leq 16$. Here, we see that the greedy algorithm is producing a near-optimal sequence, less than 27% of the size of the original de Bruijn sequence for $p \geq 4$.

To demonstrate the utility of these libraries, we validated our performance when tested against a standard experimental 10-mer library of nearly 42,000 DNA sequences for which the binding affinities of hundreds of transcription factors is known [3]. Remarkably, our library correctly recovers the high-affinity target sites, despite a nearly 4-fold reduction in library size. We were able to handle 10-mer libraries due to a 100-fold speedup in implementation over a naive one for our joker library design.

4 Conclusion

We presented a new library design that covers all k-mers with a library of size that is almost $1/|\Sigma|$ (and possibly $1/|\Sigma|^2$) smaller than current libraries, making it possible to measure interactions of significantly longer k-mers while reducing both experimental footprint and cost. We have made the implementation and library designs freely available to others.

References

1. Fordyce, P.M., Gerber, D., Tran, D., Zheng, J., Li, H., DeRisi, J.L., Quake, S.R.: De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat. Biotechnol. **28**(9), 970–975 (2010)
2. Gurobi Optimization, I.: Gurobi Optimizer Reference Manual (2015). http://www.gurobi.com
3. Hume, M.A., Barrera, L.A., Gisselbrecht, S.S., Bulyk, M.L.: UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. gku1045 (2014)

GATTACA: Lightweight Metagenomic Binning Using Kmer Counting

Victoria Popic[1], Volodymyr Kuleshov[1], Michael Snyder[2], and Serafim Batzoglou[1](✉)

[1] Department of Computer Science, Stanford University, Stanford, CA, USA
{viq,kuleshov,serafim}@stanford.edu
[2] Department of Genetics, Stanford University, Stanford, CA, USA
mpsnyder@stanford.edu

Extended Abstract

Despite their important role, microbes constitute the dark matter of the biological universe. The main limitation hindering their study is sequencing technology. The short read lengths of modern instruments – combined with various inherent difficulties associated with complex bacterial environments – make it very difficult to perform simple tasks such as accurately identifying bacterial strains, recovering their genomic sequences, and assessing their abundance. Many approaches have been proposed to address these shortcomings. Specialized library preparation techniques such as Hi-C or synthetic long reads are often very accurate, but also prohibitively complex. As a result, approaches based on contig binning are more popular in practice.

Metagenomic binning refers to the problem of grouping together partially assembled sequence fragments (or contigs) that belong to the same species. The most successful recent approaches [1, 3, 5] perform unsupervised clustering based on contig sequence composition and coverage profiles across multiple metagenomic samples. In brief, these techniques assemble de-novo bacterial contigs and estimate the coverage of each contig within each sample of a large metagenomic cohort using read mapping. This approach is accurate but has two main limitations: it requires a large cohort of samples, as well as sizable compute resources for read alignment.

In this work we present GATTACA, a lightweight framework for metagenomic binning, which (1) avoids read alignment without loss of accuracy and (2) enables efficient stand-alone analysis of single metagenomic samples. Both results are based on the finding that we can approximate contig coverages using kmer counts while still achieving the same binning accuracy as leading alignment-based methods. In addition to offering a significant speedup in coverage estimation, using kmer counts, as opposed to alignment, provides us with the exciting ability to index *offline* any publicly-available metagenomic sample. This allows us to efficiently pull in data from large growing repositories, such as the Human Microbiome Project (HMP) [4] or the EBI Metagenomics archive [2] into any metagenomic study at almost no cost. For example, our kmer count index for a typical HMP sample only requires 100MB on average. We achieve the small

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 391–392, 2017.
DOI: 10.1007/978-3-319-56970-3

space requirement by leveraging memory-efficient hashing with minimal perfect hash functions (MPHFs) and the probabilistic Bloom filter data structure. In contrast, using these datasets with read alignment would require massive downloads and expensive subsequent handling to map the reads. In terms of speedup, we found our coverage estimation time to be at least an order of magnitude faster (approximately 20×) when the index is computed offline (e.g. for recyclable public reference samples) and about 6× when the kmers are counted on-the-fly (e.g. for private samples used only once), when compared to read mapping.

While using small indices allows us to incorporate a large number of publicly-available samples into a given study, not all existing samples will improve the binning accuracy. Therefore, we propose the following two metrics for sample selection: (1) relevance and (2) diversity. More specifically, we would like to select a panel of samples which share content with the sample being analyzed (our query) but that also differ in the content that is shared. We use locality sensitive hashing and the MinHash technique, to compare the samples efficiently. At a high level, we create and index small MinHash fingerprints for each sample in the database (offline), and then extract the appropriate samples according to the fingerprint of the query.

We evaluate GATTACA for clustering contigs assembled across multiple samples (co-assemblies) and from individual samples, using both synthetic and real datasets. We compare our method to several leading alignment-based methods for metagenomic binning (such as CONCOCT [1], MetaBAT [3], and MaxBin [5]), using standardized cluster evaluation metrics and benchmarks. GATTACA was implemented in C++ and Python and is freely available at http://viq854.github.com/gattaca.

References

1. Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C.: Binning metagenomic contigs by coverage and composition. Nat. Methods **11**(11), 1144–1146 (2014)
2. Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., Jones, P., Leinonen, R., McAnulla, C., Maguire, E., et al.: Ebi metagenomics–a new resource for the analysis and archiving of metagenomic data. Nucleic Acids Res. **42**(D1), D600–D606 (2014)
3. Kang, D.D., Froula, J., Egan, R., Wang, Z.: Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ **3**, e1165 (2015)
4. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C., Knight, R., Gordon, J.I.: The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature **449**(7164), 804 (2007)
5. Wu, Y.W., Tang, Y.H., Tringe, S.G., Simmons, B.A., Singer, S.W.: Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome **2**(1), 26 (2014)

Species Tree Estimation Using ASTRAL: How Many Genes Are Enough?

Shubhanshu Shekhar[1], Sebastien Roch[2], and Siavash Mirarab[1](✉)

[1] University of California, La Jolla, San Diego, CA 92093, USA
smirarab@ucsd.edu
[2] University of Wisconsin-Madison, Madison, WI 53715, USA

Abstract. ASTRAL is a widely used method for reconstructing species trees from unrooted gene tree data. In this paper, we derive bounds on the number of gene trees needed by ASTRAL for reconstructing the true species tree with high probability. We also present some simulation results which show trends consistent with our theoretical bounds.

Keywords: Species tree estimation · Sample complexity · ASTRAL

1 Introduction

Evolutionary histories of genes and species can be discordant due to various biological processes such as incomplete lineage sorting (ILS) [1, 2]. One way to account for the discordance is to first estimate a phylogenetic tree for each gene (a gene tree) and then to summarize them to get a species tree.

ASTRAL [3] is a widely-used summary method for species tree reconstruction, and is statistically consistent under the multi-species coalescent (MSC) model [2] of ILS. ASTRAL uses dynamic programming to maximize the number of induced quartet trees shared between the species tree and the set of input gene trees, and has exact and heuristic versions. In this paper, we study ASTRAL's theoretical data requirements for successful species tree reconstruction with high probability under the MSC model and provide matching simulations results.

2 Main Results

In these results, ASTRAL* refers to the exact version of ASTRAL, f is the length of the shortest branch in the species tree in coalescent units [2], n denotes the number of leaves in the species tree and m is the number of input gene trees.

Our first result says that for small values of f, $m = \Omega(f^{-2} \log n)$ gene trees are sufficient for the correct species tree reconstruction by ASTRAL*.

The rights of this work are transferred to the extent transferable according to title 17 § 105 U.S.C.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 393–395, 2017.
DOI: 10.1007/978-3-319-56970-3

Theorem 1. *Consider a model species tree with minimum branch length f. Then, in the limit of small f and for any $\epsilon > 0$, ASTRAL* returns the true species tree with probability at least $1 - \epsilon$ if the number of input error-free gene trees satisfies*

$$m > 20 \log \left(\frac{n}{6\epsilon} \right) \frac{1}{f^2}. \tag{1}$$

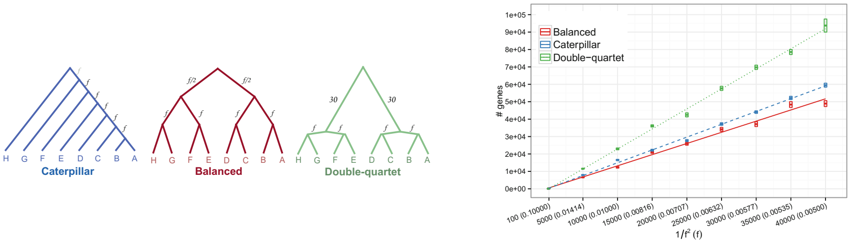

Fig. 1. Data requirement of ASTRAL-II and in simulations with $\epsilon = 0.1$. For each of the three different species tree shapes (left) and values of f (right panel; x axis), 401 replicate datasets are simulated using the MSC model, each with up to 10^5 gene trees. A binary search is used to find an approximate range for the smallest number of genes with which ASTRAL recovers the correct tree in at least 90% of the 401 replicates. Boxes show these ranges and a line is fitted to midpoints.

Our next result establishes that there exist species trees with lower bounds of error that are asymptotically similar to our upper bounds.

Theorem 2. *For any $\rho \in (0, 1)$ and $a \in (0, 1)$, there exist constants f_0 and n_0 such that the following holds. For all $n \geq n_0$ and $f \leq f_0$, there exists a species tree with n leaves and shortest branch length f such that when ASTRAL* is used with $m \leq \frac{a}{5} \frac{\log n}{f^2}$ gene trees, the event E that ASTRAL* reconstructs the wrong tree has probability*

$$P(E) \geq 1 - \rho. \tag{2}$$

These two results imply that ASTRAL* requires $\Omega(f^{-2} \log n)$ gene trees to universally guarantee correct species tree reconstruction with high probability.

Simulation results: As expected by our theoretical results, the gene tree requirement of ASTRAL increases linearly with $1/f^2$ in a simulation study (Fig. 1).

Acknowledgment. The work was supported by National Science Foundation (NSF) grant IIS-1565862 to SM and SS and NSF grant DMS-1149312 (CAREER) to SR.

References

1. Maddison, W.P.: Gene trees in species trees. Syst. Biol. **46**(3), 523–536 (1997)
2. Degnan, J.H., Rosenberg, N.A.: Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. **24**(6), 332–340 (2009)
3. Mirarab, S., Warnow, T.: ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics **31**(12), i44–i52 (2015)

Reconstructing Antibody Repertoires from Error-Prone Immunosequencing Datasets

Alexander Shlemov[1], Sergey Bankevich[1], Andrey Bzikadze[1],
Yana Safonova[1(✉)], and Pavel A. Pevzner[1,2]

[1] Center for Algorithmic Biotechnology, Institute of Translational Biomedicine,
St. Petersburg State University, Saint Petersburg, Russia
safonova.yana@gmail.com

[2] Dept. of Computer Science and Engineering, University of California,
San Diego, La Jolla, CA, USA

1 Introduction

Recent progress in sequencing technologies enabled generation of high-throughput full-length antibody sequences using read-pairs formed by overlapping reads within read-pairs. However, transforming error-prone Rep-seq datasets into accurate *antibody repertoires* is a challenging bioinformatics problem [1, 2, 4] that is a prerequisite for a multitude of downstream studies of adaptive immune system.

Until 2013, there were few attempts to develop algorithms for full-length antibody repertoire reconstruction since it was unclear how to derive accurate repertoires from error-prone reads produced by the low-throughput 454 sequencing technology. However, in the last three years, immunology laboratories developed various Rep-seq protocols aimed at generating high-throughput Rep-seq datasets and constructing repertoires based on more accurate Illumina MiSeq reads.

Recently, several tools for constructing full-length antibody repertoires were developed, including MiGEC [3], pRESTO [4], MiXCR [1], and IgRepertoireConstructor [2]. Some of these tools also are able to utilize information about molecular barcodes. However, quality assessment of the constructed antibody repertoire and, thus, benchmarking of various repertoire construction algorithms are still poorly addressed problems.

In this paper, we use barcoded Rep-seq datasets and simulated antibody repertoires to benchmark various repertoire construction algorithms. Our novel toolkit includes IgReC, a tool for antibody repertoire construction from both barcoded and non-barcoded immunosequencing data, and IgQUAST, a tool for quality assessment of antibody repertoires. IgReC package is freely available at http://yana-safonova.github.io/ig_repertoire_constructor.

Alexander Shlemov and Sergey Bankevich—These authors contributed equally.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 396–397, 2017.
DOI: 10.1007/978-3-319-56970-3

2 Discussion

Our benchmarking on non-barcoded data revealed that there is still no single repertoire construction tool that works better than others across the diverse types of Rep-seq datasets. However, IgReC is currently a tool of choice for analyzing hypermutated repertoires.

We also compared IgReC in a blind mode against pRESTO and MiGEC that utilize information about molecular barcodes. Benchmarking on simulated barcoded datasets revealed that while all tools result in high sensitivity, their precision varies and becomes rather low in the case of high PCR error rates. Surprisignly, repertoires reported by IgReC tool (in the blind mode) improved on the repertoires constructed by the specialized tools that use barcoding information.

Acknowledgements. We are indebted to Dmitry Chudakov, Dmitry Bolotin, Mikhail Shugay, Jason Vander Heiden, and Steven Kleinstein for productive discussions and assistance in benchmarking MiXCR, MiGEC, and pRESTO tools.

Funding. This project is supported by Russian Science Foundation (grant No 14-50-00069).

References

1. Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., Chudakov, D.M.: MiXCR: software for comprehensive adaptive immunity proling. Nat. Methods **12**(5), 3801 (2015)
2. Safonova, Y., Bonissone, S., Kurpilyansky, E., Starostina, E., Lapidus, A., Stinson, J., DePalatis, L., Sandoval, W., Lill, J., Pevzner, P.A.: IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. Bioinformatics **31**(12), i53–61 (2015)
3. Shugay, M., Britanova, O., Merzlyak, E., Turchaninova, M., Mamedov, I., Tuganbaev, T., Bolotin, D., Staroverov, D., Putintseva, E., Plevova, K., Linnemann, C., Shagin, D., Pospisilova, S., Lukyanov, S., Schumacher, T., Chudakov, D.M.: Towards error-free proling of immune repertoires. Nat Methods **11**, 6535 (2014)
4. Vander Heiden, J.A., Yaari, G., Uduman, M., Stern, J.N., O'Connor, K.C., Haer, D.A., Vigneault, F., Kleinstein, S.H.: pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics **30**(13), 1930–1932 (2014)

NetREX: Network Rewiring Using EXpression - Towards Context Specific Regulatory Networks

Yijie Wang[1], Dong-Yeon Cho[1], Hangnoh Lee[2], Brian Oliver[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] Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, 50 South Drive,
Bethesda, MD 20892, USA

Extended Abstract

Understanding gene regulation is a fundamental step towards understanding of how cells function and respond to environmental cues and perturbations. An important step in this direction is the ability to infer the transcription factor (TF)-gene regulatory network (GRN). However gene regulatory networks are typically constructed disregarding the fact that regulatory programs are conditioned on tissue type, developmental stage, sex, and other factors. Due to lack of the biological context specificity, these context-agnostic networks may not provide insight for revealing the precise actions of genes for a specific biological system under concern. Collecting multitude of features required for a reliable construction of GRNs such as physical features (TF binding, chromatin accessibility) and functional features (correlation of expression or chromatin patterns) for every context of interest is costly. Therefore we need methods that are able to utilize the knowledge about a context-agnostic network (or a network constructed in a related context) for construction of a context specific regulatory network.

To address this challenge we developed a computational approach that utilizes expression data obtained in a specific biological context and a GRN constructed in a different but related context to construct a context specific GRN. Our method, NetREX, is inspired by network component analysis that estimates TF activities and their influences on target genes given predetermined topology of a TF-gene regulatory network. To predict a network under a different condition, NetREX removes the restriction that the topology of the TF-gene regulatory network is fixed and allows for adding and removing edges to that network. Mathematically, we use ℓ_0 norm to directly handle the number of removed and newly added edges as well as induce sparse solutions in our formulation.

Yijie Wang and Dong-Yeon Cho — Equal contribution.

The rights of this work are transferred to the extent transferable according to title 17 §105 U.S.C.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 398–399, 2017.
DOI: 10.1007/978-3-319-56970-3

Unlike the widely used strategy, which is replacing the non-convex ℓ_0 norm by its convex relaxation ℓ_1 norm, we focus on the harder problem involving ℓ_0 norm and provide a number of rigorous derivations and results allowing us to adopt the recently proposed Proximal Alternative Linearized Maximization (PALM) algorithm. In addition, we also proved the convergence of the NetREX algorithm.

We tested our NetREX on simulated data and found that NetREX is able to dramatically improve the accuracy of the regulatory networks as long as the prior network and the gene expression are not very noisy. Subsequently, we applied NetREX for constructing regulatory networks for adult female flies. We used the network constructed in a recent study as the prior network, which was build by integrating diverse data sets including TF binding, evolutionarily conserved sequence motifs and so on. Starting with this network, we utilized a new expression data set that we collected for adult female flies where perturbations in expression were achieved by genetic deletions. We accessed the biological relevance of the predicted networks by using Gene Ontology annotations and physical protein-protein interactions. The networks predicted by NetREX showed higher biological consistency than alternative approaches. In addition, we used the list of recently identified targets of the Doublesex (DSX) transcription factor to demonstrate the predictive power of our method.

E Pluribus Unum: United States of Single Cells

Joshua D. Welch[1(✉)], Alexander Hartemink[2], and Jan F. Prins[1]

[1] Department of Computer Science, The University of North Carolina,
Chapel Hill, USA
{jwelch,prins}@cs.unc.edu
[2] Department of Computer Science, Duke University, Durham, USA

Extended Abstract

Single cell genomic techniques promise to yield key insights into the dynamic interplay between gene expression and epigenetic modification. However, the experimental difficulty of performing multiple measurements on the same cell currently limits efforts to combine multiple genomic data sets into a united picture of single cell variation [1, 2]. The current understanding of epigenetic regulation suggests that any large changes in gene expression, such as those that occur during differentiation, are accompanied by epigenetic changes. This means that if cells undergoing a common process are sequenced using multiple genomic techniques, examining any of the genomic quantities should reveal the same underlying biological process. For example, the main difference among cells undergoing differentiation will be the extent of their differentiation progress, whether you look at the gene expression profiles or the chromatin accessibility profiles of the cells.

We reasoned that this property of single cell data could be used to infer correspondence between different types of genomic data. To infer single cell correspondences, we use a technique called manifold alignment. Intuitively, manifold alignment constructs a low-dimensional representation (manifold) for each of the observed data types, then projects these representations into a common space (alignment) in which measurements of different types are directly comparable [3, 4]. To the best of our knowledge, manifold alignment has never been used in genomics. However, other application areas recognize the technique as a powerful tool for multimodal data fusion, such as retrieving images based on a text description, and multilingual search without direct translation [4].

We show for the first time that it is possible to construct cell trajectories, reflecting the changes that occur in a sequential biological process, from single cell epigenetic data. In addition, we present an approach called MATCHER that computationally circumvents the experimental difficulties of performing multiple genomic measurements on a single cell by inferring correspondence between single cell transcriptomic and epigenetic measurements performed on different cells of the same type. MATCHER works by first learning a separate manifold for the trajectory of each kind of genomic data, then aligning the manifolds to infer a shared trajectory in which cells measured using different techniques are directly comparable. Because there is, in general, no actual cell-to-cell correspondence

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 400–401, 2017.
DOI: 10.1007/978-3-319-56970-3

between datasets measured with different experimental techniques, MATCHER *generates* corresponding measurements by predicting what each type of measurement *would* look like at a given point in the process. Using scM&T-seq data, we confirm that MATCHER accurately predicts true single cell correlations between DNA methylation and gene expression without using known cell correspondence information.

We also downloaded publicly available single cell genomic data from a total of 4,974 single mouse embryonic stem cells grown in serum. Each cell in this dataset was individually assayed using one of four experimental techniques: RNA-seq, scM&T-seq, ATAC-seq, or ChIP-seq. We used MATCHER to infer correlations among these four measurements. This analysis gave novel insights into the changes that cells undergo as they transition from pluripotency to a differentiation primed state.

We found three main results. First, chromatin accessibility and histone modification changes largely fall into two anti-correlated categories: silencing of pluripotency factor binding sites and repression of lineage-specific genes by chromatin remodeling factors. Second, the action of pluripotency transcription factors is gradually removed by both transcriptional silencing of the genes and epigenetic silencing of the binding sites for these factors. In contrast, regulation of chromatin remodeling factor activity occurs primarily at the epigenetic level, largely unaccompanied by changes in the expression of the chromatin remodeling factors. Third, DNA methylation changes are strongly coupled to gene expression changes early in the process of differentiation priming, but the degree of coupling drops sharply later in the process.

Our work is a first step toward a united picture of heterogeneous transcriptomic and epigenetic states in single cells. MATCHER promises to be a powerful tool as single cell genomic approaches continue to generate revolutionary discoveries in fields ranging from cancer biology and regenerative medicine to developmental biology and neuroscience.

References

1. Bock, C., Farlik, M., Sheffield, N.C.: Multi-omics of single cells: strategies and applications. Trends Biotechnol. **34**(8), 605608 (2016)
2. Macaulay, I.C., Ponting, C.P., Voet, T.: Single-cell multiomics: multiple measurements from single cells. Trends Genet. **33**(2), 155–168 (2017)
3. Ham, J., Lee, D.D., Saul, L.K.: Semisupervised alignment of manifolds. In: AISTATS, p. 120127 (2005)
4. Wang, C., Mahadevan, S.: A general framework for manifold alignment. In: AAAI (2009)

ROSE: A Deep Learning Based Framework for Predicting Ribosome Stalling

Sai Zhang[1], Hailin Hu[2], Jingtian Zhou[2], Xuan He[1], Tao Jiang[3,4,5], and Jianyang Zeng[1](✉)

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
zengjy321@tsinghua.edu.cn
[2] School of Medicine, Tsinghua University, Beijing, China
[3] Department of Computer Science and Engineering, University of California, Riverside, CA, USA
[4] MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST/Department of Computer Science and Technology, Tsinghua University, Beijing, China
[5] Institute of Integrative Genome Biology, University of California, Riverside, CA, USA

Abstract. Translation elongation plays a crucial role in multiple aspects of protein biogenesis, e.g., differential expression, cotranslational folding and secretion. However, our current understanding on the regulatory mechanisms underlying translation elongation dynamics and the functional roles of ribosome stalling in protein synthesis still remains largely limited. Here, we present a deep learning based framework, called ROSE, to effectively predict ribosome stalling events in translation elongation from coding sequences. Our validation results on both human and yeast datasets demonstrate superior performance of ROSE over conventional prediction models. With high prediction accuracy and robustness across different datasets, ROSE shall provide an effective index to estimate the translational pause tendency at codon resolution. We also show that the ribosome stalling score (RSS) output by ROSE correlates with diverse putative regulatory factors of ribosome stalling, e.g., codon usage bias, codon cooccurrence bias, proline codons and N^6-methyladenosine (m^6A) modification, which validates the physiological relevance of our approach. In addition, our comprehensive genome-wide *in silico* studies of ribosome stalling based on ROSE recover several notable functional interplays between elongation dynamics and cotranslational events in protein biogenesis, including protein targeting by the signal recognition particle (SRP) and protein secondary structure formation. Furthermore, our intergenic analysis suggests that the enriched ribosome stalling events at the 5' ends of coding sequences may be involved in the modulation of translation efficiency. These findings indicate that ROSE can provide a useful index to estimate the probability of ribosome stalling and offer a powerful tool to analyze the large-scale ribosome profiling data, which will further expand our understanding on translation elongation dynamics. The full version of this work can be found as a preprint at https://doi.org/10.1101/067108.

S. Zhang, H. Hu, J. Zhou — These authors contributed equally to this work.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 402–403, 2017.
DOI: 10.1007/978-3-319-56970-3

Acknowledgements. This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300 and 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003 and 61472205, the US National Science Foundation Grant DBI-1262107 and IIS-1646333, the China's Youth 1000-Talent Program, and the Beijing Advanced Innovation Center for Structural Biology.

Author Index