Lecture Notes in Bioinformatics 10330

Subseries of Lecture Notes in Computer Science

More information about this series at http://www.springer.com/series/5381

Zhipeng Cai · Ovidiu Daescu
Min Li (Eds.)

Bioinformatics Research and Applications

13th International Symposium, ISBRA 2017 Honolulu, HI, USA, May 29 – June 2, 2017 Proceedings

Springer

Editors
Zhipeng Cai
Georgia State University
Atlanta, GA
USA

Min Li
Central South University
Changsha
China

Ovidiu Daescu
University of Texas at Dallas
Richardson, TX
USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Bioinformatics
ISBN 978-3-319-59574-0 ISBN 978-3-319-59575-7 (eBook)
DOI 10.1007/978-3-319-59575-7

Library of Congress Control Number: 2017941549

LNCS Sublibrary: SL8 – Bioinformatics

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

On behalf of the Program Committee, we would like to welcome you to the proceedings of the 13th edition of the International Symposium on Bioinformatics Research and Applications (ISBRA 2017), held in Honolulu, Hawaii, May 29 to June 2, 2017. The symposium provides a forum for the exchange of ideas and results among researchers, developers, and practitioners working on all aspects of bioinformatics and computational biology and their applications. This year we received 118 submissions in response to the call for extended abstracts. The Program Committee decided to accept 27 of them for full publication in the proceedings and oral presentation at the symposium. We also accepted 24 of them for oral presentation and short abstract publication in the proceedings. Furthermore, we also received 18 submissions in response to the call for short abstracts.

The technical program invited keynote talks by Prof. Michael Q. Zhang from The University of Texas at Dallas and Tsinghua University. Prof. Zhang reviewed the history of computational genome regulation and then introduced some new biochemical (BL-Hi-C), biophysical (super-resolution imaging), and bioinformatics (MICC, 3CPET, FIND) technology developments that may be used for studying 3D genomes and disease markers in the near future.

We would like to thank the Program Committee members and the additional reviewers for volunteering their time to review and discuss symposium papers. We would like to extend special thanks to the steering and general chairs of the symposium for their leadership, and to the finance, publicity, workshops, local organization, and publications chairs for their hard work in making ISBRA 2017 a successful event. Last but not least we would like to thank all authors for presenting their work at the symposium.

April 2017

Zhipeng Cai
Ovidiu Daescu
Min Li

Organization

Steering Committee

Dan Gusfield	University of California Davis, USA
Ion Mandoiu	University of Connecticut, USA
Yi Pan (Chair)	Georgia State University, USA
Marie-France Sagot	INRIA, France
Ying Xu	University of Georgia, USA
Alexander Zelikovsky	Georgia State University, USA

General Chair

Alexander Zelikovsky	Georgia State University, USA

Program Chairs

Zhipeng Cai	Georgia State University, USA
Ovidiu Daescu	The University of Texas at Dallas, USA
Min Li	Central South University, China

Finance Chair

Anu G. Bourgeois	Georgia State University, USA

Publications Chair

Pavel Skums	Georgia State University, USA

Publicity Chairs

Chunyu Ai	University of South Carolina Upstate, USA
Shaoliang Peng	National University of Defense Technology, China
Xiang Wan	Hong Kong Baptist University, China
Gangman Yi	Dongguk University, Korea

Workshop Chairs

Yaohang Li	Old Dominion University, USA
Anu G. Bourgeois	Georgia State University, USA
Wooyoung Kim	University of Washington Bothell, USA

Award Chair

Raj Sunderraman	Georgia State University, USA

Publication Chair

Pavel Skums	Georgia State University, USA

Webmasters

Igor Mandric	Georgia State University, USA
Sergey Knyazev	Georgia State University, USA

Program Committee

Kamal Al Nasr	Tennessee State University, USA
Max Alekseyev	George Washington University, USA
Mukul S. Bansal	University of Connecticut, USA
Robert Beiko	Dalhousie University, Canada
Paola Bonizzoni	Università di Milano-Bicocca, Italy
Zhipeng Cai	Georgia State University
Doina Caragea	Kansas State University, USA
Xing Chen	National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, China
Ovidiu Daescu	University of Texas at Dallas, USA
Fei Deng	University of California, Davis
Lei Deng	Central South University
Oliver Eulenstein	Iowa State University, USA
Lin Gao	Xidian University, China
Olga Glebova	Georgia State University
Jiong Guo	Shandong University, China
Xuan Guo	Oak Ridge National Laboratory, USA
Jieyue He	Southeast University
Zengyou He	Hong Kong University of Science and Technology, SAR China
Steffen Heber	NCSU
Xing Hua	
Jinling Huang	East Carolina University, USA
Mingon Kang	Kennesaw State University, USA
Wooyoung Kim	University of Washington Bothell, USA
Danny Krizanc	Wesleyan University, USA
Xiujuan Lei	Shaanxi Normal University, China
Jing Li	Case Western Reserve University, USA
Min Li	Central South University
Shuai Cheng Li	City University of Hong Kong, SAR China
Yaohang Li	Old Dominion University, USA
Yingshu Li	Georgia State University, Atlanta, USA

Xiaowen Liu	Indiana University-Purdue University Indianapolis, USA
Ion Mandoiu	University of Connecticut, USA
Fenglou Mao	National Institute of Health, USA
Giri Narasimhan	Florida International University
Chongle Pan	Oak Ridge National Laboratory, USA
Steven Pascal	Old Dominion University, USA
Andrei Paun	University of Bucharest, Romania
Nadia Pisanti	Universita di Pisa, Italy and Erable Team, Italy; Inria, France
Mukul S. Bansal	University of Connecticut, USA
Russell Schwartz	Carnegie Mellon University, USA
Joao Setubal	University of São Paulo, Brazil
Xinghua Shi	University of North Carolina at Charlotte, USA
Yi Shi	Shanghai Jiao Tong University, China
Pavel Skums	Georgia State University, USA
Ileana Streinu	Smith College, Northampton, USA
Chia-Yu Su	Taipei Medical University, Taiwan
Wing-Kin Sung	National University of Singapore, Singapore
Sing-Hoi Sze	Texas A&M University, USA
Weitian Tong	Georgia Southern University, USA
Gabriel Valiente	Technical University of Catalonia, Spain
Xiang Wan	HKBU
Jianxin Wang	Central South University
Li-San Wang	University of Pennsylvania
Lusheng Wang	City University of Hong Kong, SAR China
Peng Wang	Shanghai Advanced Research Institute, Chinese Academy of Science, China
Seth Weinberg	Virginia Commonwealth University, USA
Fangxiang Wu	University of Saskatchewan, Canada
Yufeng Wu	University of Connecticut, USA
Zeng Xiangxiang	
Xiuchun Xiao	Fudan University, China
Minzhu Xie	Hunan Normal University, China
Dechang Xu	Harbin Institute of Technology, China
Can Yang	HKBU
Ashraf Yaseen	Texas A&M University, Kingsville, USA
Guoxian Yu	Southwest University
Ning Yu	Georgia State University
Alex Zelikovsky	Georgia State University, USA
Chi Zhang	Indiana University, USA
Fa Zhang	Institute of Computing Technology
Le Zhang	Southwest University
Xue Zhang	Tufts University, USA
Yanqing Zhang	Georgia State University, USA
Leming Zhou	University of Pittsburgh
Quan Zou	Tianjin University, China

Additional Reviewers

Abdelrasoul, Maha
Aldabagh, Hind
Alexeev, Nikita
Antipov, Dmitry
Artyomenko, Alexander
Arunachalam, Harish Babu
Avdeyev, Pavel
Biswas, Abhishek
Chen, Wei
Chu, Chong
Daescu, Kelly
Della Vedova, Gianluca
Diaz Tula, Antonio
Elhefnawy, Wessam
Farhana, Effat
Frith, Martin
Glebova, Olga
He, Jing
Hu, Jialu
Hu, Xiaoming
Hu, Xihao
Icer, Pelin
Ionescu, Vlad
Knyazev, Sergey
Lan, Wei
Li, Jin
Li, Leon
Li, Xin
Liu, Bin
Llabrés, Mercè
Mandric, Igor
Melnyk, Andrii
Moon, Jucheol
Muntean, Radu
Olariu, Ciprian
Patterson, Murray
Pei, Jingwen
Peng, Xiaoqing
Perkins, Patrick
Ren, Xianwen
Rizzi, Raffaella
Sheng, Tao
Shi, Jian-Yu
Sun, Yazhou
Trivette, Andrew
Vyatkina, Kira
Wan, Changlin
Wu, Hao
Wu, Yue
Xiangxiang, Zeng
Yang, Frank
Yuan, Xiguo
Zaccaria, Simone
Zhao, Junfei
Zhao, Qi
Zhu, Shanfeng
Zhu, Zexuan

Abstract of Invited Papers

Copy Number Aberration Based Cancer Type Prediction with Convolutional Neural Networks

Yuchen Yuan[1,2], Yi Shi[2], Xianbin Su[2], Xin Zou[2], Qing Luo[2], Weidong Cai[1], Zeguang Han[2], and David Dagan Feng[1]

[1] School of Information Technologies,
The University of Sydney, Sydney, NSW 2008, Australia
{yuchen.yuan,tom.cai,dagan.feng}@sydney.edu.au
[2] Key Laboratory of Systems Biomedicine,
Shanghai Center for Systems Biomedicine,
Shanghai Jiaotong University, Shanghai 200240, China
{yishi,xbsu,x.zou,simonluo,hanzg}@sjtu.edu.cn

Abstract. Cancer is a category of disease that causes abnormal cell growths and immortality. It usually incarnates into tumor form that potentially invade or metastasize to remote parts of human body [1]. During the past decade, with the developments of DNA sequencing technology, large amounts of sequencing data have become available which provides unprecedented opportunities for advanced association studies between somatic mutations and cancer types/subtypes [2–7], which may contribute to more accurate somatic mutation based cancer typing (SMCT). In existing SMCT methods however, the absence of feature quantification and high-level feature extraction is a major obstacle in improving the classification performance. To address this issue, we propose DeepCNA, an advanced convolutional neural network (CNN) based classifier, which utilizes copy number aberrations (CNAs) [8–10] and HiC data [11] for cancer typing. DeepCNA consists of two steps: firstly, the CNA data is pre-processed by clipping, zero padding and reshaping; secondly, the processed data is fed into a CNN classifier, which extracts high-level features for accurate classification [12].

We conduct experiments on the newly proposed COSMIC CNA dataset, which contains 25 types of cancer. Controlled variable experiments indicate that the 2D CNN with both cell lines of HiC data (hESC and IMR90) contributes to the optimal performance. We then compare DeepCNA with three widely adopted data classifiers, the results of which exhibit the remarkable advantages of DeepCNA, which has achieved significant performance improvements in terms of testing accuracy (78%) against the comparison methods. We have demonstrated the advantages and potentials of the DeepCNA model for somatic point mutation based gene data processing, and suggest that the model can be extended and transferred to other complex genotype-phenotype association studies, which we believe will benefit many related areas [13, 14].

Yuchen Yuan, Yi Shi—These authors contribute equally as co-first authors.

References

1. Feuerstein, M.: Defining cancer survivorship. J. Cancer Survivorship **1**(1), 5–7, (2007)
2. Yang, K., Li, J., Cai, Z., Lin, G.: A model-free and stable gene selection in microarray data analysis. In: IEEE Symposium of BioInformatics BioEngineering (BIBE), Minneapolis, MN, USA, pp. 3–10 (2005)
3. Yang, K., Cai, Z., Li, J., Lin, G.: A stable gene selection in microarray data analysis. BMC Bioinform. **7**(1), 228 (2006)
4. Cai, Z., Goebel, R., Salavatipour, M.R., Lin, G.: Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. BMC Bioinform. **8**(1), 206 (2007)
5. Cai, Z., Xu, L., Shi, Y., Salavatipour, M.R., Goebel, R., Lin, G.: Using gene clustering to identify discriminatory genes with higher classification accuracy. In: IEEE Symposium of BioInformatics BioEngineering (BIBE), Arlington, VA, USA, pp. 235–242 (2006)
6. Cai, Z., Zhang, T., Wan, X.-F.: A computational framework for influenza antigenic cartography. PLoS Comput. Biol. **6**(10), e1000949 (2010)
7. Cai, Z., Ducatez, M.F., Yang, J., Zhang, T., Long, L., Boon, A.C., Webby, R.J., Wan, X.: Identifying antigenicity associated sites in highly pathogenic H5N1 Influenza Virus Hemagglutinin by using sparse learning. J. Mol. Biol. **422**(1), 145–155 (2012)
8. Bakhoum, S.F., Swanton, C.: Chromosomal instability, aneuploidy, and cancer. Frontiers Oncol. **4**, 161 (2014)
9. Burrell, R.A., McGranahan, N., Bartek, J., Swanton, C.: The causes and consequences of genetic heterogeneity in cancer evolution. Nature **501**(7467), 338–345 (2013)
10. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., et al.: Pan-cancer patterns of somatic copy number alteration. Nat. Genet. **45**(10), 1134–1140 (2013)
11. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**(7398), 376–380 (2012)
12. Yuan, Y., Shi, Y., Li, C., Kim, J., Cai, W., Han, Z., et al.: DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinform. **17**(17), 243 (2016)
13. Longo, D.L.: Tumor heterogeneity and personalized medicine. N. Engl. J. Med. **366**(10), 956–957 (2012)
14. Franken, B., de Groot, M.R., Mastboom, W.J., Vermes, I., van der Palen, J., Tibbe, A.G., et al.: Circulating tumor cells, disease recurrence and survival in newly diagnosed breast cancer. Breast Cancer Res. **14**(5), 1–8 (2012)

Predicting Human Microbe-Disease Associations via Binary Matrix Completion

Jian-Yu Shi[1], Hua Huang[2], Yan-Ning Zhang[3], and Siu-Ming Yiu[4]

[1] School of Life Sciences, Northwestern Polytechnical University, Xi'an, China
jianyushi@nwpu.edu.cn
[2] School of Software and Microelectronics,
Northwestern Polytechnical University, Xi'an, China
1363351294@qq.com
[3] School of Computer Science,
Northwestern Polytechnical University, Xi'an, China
ynzhang@nwpu.edu.cn
[4] Department of Computer Science, the University of Hong Kong,
Pok Fu Lam, Hong Kong
smyiu@cs.hku.hk

With the help of sequencing techniques (e.g. 16S ribosomal RNA sequencing) [1], Human Microbiome Project has revealed that there are diverse communities of microbes in a human intestine, which provides a nutrient-rich and temperature-fixed habitat for microbes. The sequential works have observed that there exists a significant mutual influence between microbes and their host. It is surprising that except for conventional infectious diseases, a wide range of noninfectious diseases is closely associated with microbes, such as cancer, obesity [2], diabetes, kidney stones and systemic inflammatory response syndrome. On the one side, the tremendous amount of microbiome genes and their products can lead a diverse range of biological activities, which serve as a physiological complement in their host body in a wide range, involving metabolic capabilities, pathogens, immune system, and gastrointestinal development [3]. On the other side, the microbes can be greatly influenced by their dynamic habitat in the human body, which undergoes frequent changes caused by diverse environmental variables, such as season, host diet, smoking, hygiene and use of antibiotics. Thus, this mutual association between the host and its microbiota can further modify transcriptomic, proteomic and metabolic profiles of the human host. However, the identification of microbe-noninfectious disease associations (MDAs) requires time-consuming and costly experiments and always bears the limitation of microbe cultivation. Even worse, many bacteria cannot be cultivated at all by current culturing bio-techniques. Fortunately, the number of MDAs found in both experiments and clinic is growing. For example, Ma et al. published the first MDA database, Human Microbe-Disease Association Database (HMDAD) recently, by collecting a large number of MDAs from previously published literature [4]. The growing number of MDAs enables us to perform a systematic analysis, discovery and understanding on the mechanism of microbe-related non-infectious diseases in a new insight. As one of the most important steps to achieve that goal, the discovery or prediction of potential MDAs provides an approach to understand the mechanism of non-infectious disease

formation and development and develop novel methods for disease diagnosis and therapy. As the promising complement of experiment-based approaches, computational approaches, especially machine learning-based approaches, are able to predict MDA candidates among a large number of microbe-disease pairs. They cannot only reduce the cost and time of relevant experiments, but also output the candidates, of which even though the involving microbes cannot cultured. Nevertheless, a few of efforts have been made to develop computational models for MDA prediction on a large scale. Very recently, a pioneering work constructing an MDA network based on HMDAD develops an approach KATZHMDA for predicting potential MDAs [5]. KATZHMDA regards the prediction of MDS as link prediction on the constructed MDA network. In this work, we first model MDA prediction as a problem of matrix completion (Fig. 1), then propose a new approach based on Binary Matrix Completion (BMCMDA) to predict potential MDAs. BMCMDA is able to predict new MDAs on a large scale, by only using known microbe-disease association network. Its performance is evaluated by both leave-one-out cross validation (LOOCV) and 5-fold cross validation (5-CV) on HMDAD database, where the whole procedure of 5-CV was repeated 100 times and both the mean and the standard deviation of predicting performance over 100 rounds of 5-CVs were recorded. Finally, in terms of Area Under Receiver-Operating Characteristics, BMCMDA achieves 0.9049 in LOOCV and 0.8954 ± 0.0034 in 5CV, while the state-of-the-art KATZHMDA only achieves 0.8382 and 0.8301 ± 0.0033 respectively. The significantly outperformed prediction achieved by BMCMAD demonstrates its superiority for predicting microbe-disease associations on a large scale.

Acknowledgments. This work was supported by RGC Collaborative Research Fund (CRF) of Hong Kong (C1008-16G), National High Technology Research and Development Program of China (No. 2015AA016008), the Fundamental Research Funds for the Central Universities of China (No. 3102015ZY081), the Program of Peak Experience of NWPU (2016) and partially supported by the National Natural Science Foundation of China (No. 61473232, 91430111).

References

1. Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A., Creasy, H.H., Earl, A.M., Fitzgerald, M., Fulton, R.S.: Structure, function and diversity of the healthy human microbiome. Nature **486**, 207–214 (2012)
2. Zhang, H., Dibaise, J.K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., Parameswaran, P., Crowell, M.D., Wing, R.A., Rittmann, B.E.: Human gut microbiota in obesity and after gastric bypass. Proc. Nat. Acad. Sci. U.S.A. **106**, 2365–2370 (2009)
3. Ventura, M., O'Flaherty, S., Claesson, M.J., Turroni, F., Klaenhammer, T.R., Van, S.D., O'Toole, P.W.: Genome-scale analyses of health-promoting bacteria: probiogenomics. Nat. Rev. Microbiol. **7**, 61–72 (2009)
4. Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., Yang, J., Kong, W., Zhou, X., Cui, Q.: An analysis of human microbe–disease associations. Briefings Bioinform. **18**, 85–97 (2017)
5. Chen, X., Huang, Y.A., You, Z.H., Yan, G.Y., Wang, X.S.: A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics **33**, 733–739 (2017)

Characterization of Kinase Gene Expression and Splicing Profile in Prostate Cancer with RNA-Seq Data

Huijuan Feng[1], Tingting Li[2], and Xuegong Zhang[1,3]

[1] MOE Key Laboratory of Bioinformatics,
Bioinformatics Division/Center for Synthetic and Systems Biology,
TNLIST and Department of Automation, Tsinghua University,
Beijing 100084, China
fhj11@mails.tsinghua.edu.cn
[2] Department of Biomedical Informatics, Institute of Systems Biomedicine,
School of Basic Medical Sciences,
Peking University Health Science Center, Beijing 100191, China
litt@hsc.pku.edu.cn
[3] School of Life Sciences, Tsinghua University, Beijing 100084, China
zhangxg@tsinghua.edu.cn

Abstract. Alternative splicing is a ubiquitous post-transcriptional process in most eukaryotic genes. Aberrant splicing isoforms and abnormal isoform ratios can contribute to cancer development. Kinase genes are key regulators of many cellular processes. Multiple kinases are found to be oncogenic. RNA-Seq provides a powerful technology for genome-wide study of alternative splicing. But this potential has not been fully demonstrated on cancers yet. We characterized the transcriptome profile of prostate cancer using RNA-Seq data on both differential expression and differential splicing, with an emphasis on kinase genes and their splicing variations. We identified distinct gene groups from differential expression and splicing analysis, which suggested that alternative splicing adds another level to gene regulation in cancer. Enriched GO terms of differentially expressed and spliced kinase genes were found to play different roles in regulation of cellular metabolism. Function analysis showed that differentially spliced exons of these genes are significantly enriched in protein kinase domains. Among them, we found that gene CDK5 has isoform switching between prostate cancer and benign tissues, which may affect cancer development by changing androgen receptor (AR) phosphorylation. The observation was validated in another RNA-Seq dataset of prostate cancer cell lines. Our work brings new understanding to the role of alternatively spliced kinases in prostate cancer and demonstrates the use of RNA-Seq data in studying alternative splicing in cancer.

Keywords: Prostate cancer · Alternative splicing · Kinase · CDK5 · Isoform switching

Identifying Conserved Protein Complexes Across Multiple Species via Network Alignment

Bo Song[1], Jianliang Gao[1,2], Xiaohua Hu[1], Yu Sheng[2], and Jianxin Wang[2]

[1] College of Computing and Informatics, Drexel University, Philadelphia, USA
[2] School of Information Science and Engineering, Central South University, Changsha, China
gaojianliang@csu.edu.cn

A protein complex is a bimolecular that contains a number of proteins interacting with each other to perform different cellular functions [1]. The identification of protein complexes in a protein-protein interaction (PPI) network [2] can, therefore, lead to a better understanding of the roles of such a network in different cellular systems. The protein complex identification problem has received a lot of attentions, and a considerable number of techniques have been proposed to address such problem. By representing a PPI network as a graph [3], whose vertices represent proteins and edges as interactions between proteins, these algorithms are able to identify clusters in single PPI network based on different graph properties [4]. For example, an uncertain graph model based method is proposed to detect protein complex from a PPI network [5]. However, they focused on finding protein complexes in a single PPI network, and finding conserved protein complexes from multiple PPI networks still remain challenging.

In this paper, we identify the problem of finding conserved protein complexes via aligning multiple PPI networks. In this way, the knowledge of protein complexes in well-studied species can be extended to that of poor-studied species. Then, we propose an efficient method to find conserved protein complexes from multiple PPI networks. By taking the feature of subnetwork connectivity into consideration, the proposed method improves the coverage significantly without compromising of the consistency in the aligned results.

Given the multiple PPI networks $(G_1, G_2, \ldots, G_\xi)$ and target protein complex M_0 from the target PPI network G_t, the alignment process mainly includes:

(1) Generate initial candidate pools. Only those proteins that have links with given protein complex can be selected as candidate proteins since links represent the biological similarity between proteins across PPI networks. For each aligned network G_i, $1 \leq i \leq \xi$, we construct a pool for a given protein complex M_0, where $M_0 \in G_t$. Every vertex $v \in G_i$ is put into the pool of G_i if it has link with any vertex in M_0. Then, the initial subnetworks M are selected randomly from the pools.

(2) Optimal determination by simulated annealing. Simulated annealing process adopts iteration method for global optimal solution. In each loop, a protein from the candidate pool is chosen randomly to be determined as aligned protein in the

corresponding PPI network. There are two kinds of proteins that are possible to be moved out from the current alignment solution. The first kind is the protein whose score is the lowest in the current solution. The other kind is the protein whose corresponding vertex in the current subnetwork is not connected with other vertices, i.e., its degree is zero. If the new candidate solution achieves higher score, it will take place the previous solution. If not, it still has chance to replace the prior solution with a probability of $(rand(0,1) < e^{\frac{\Delta\Phi}{T_i}})$, where $\Delta\Phi$ is the amount of change score, T_i is the temperature of simulated annealing. Finally, the algorithm returns the best solution as the alignment of protein complexes $M = \{M_1, M_2, \ldots, M_\xi\}$. Overall, we utilize both the biological similarity between proteins and the topological structure to assign scores on subnetworks for simulated annealing process. Formally, given a protein complex of target network $M_0 \subseteq G_t$, its match result $\{M_1, M_2, \ldots, M_\xi\}$ in aligned networks, where $M_k \subseteq G_k$, is assigned a real-valued score Φ:

$$\Phi = \sum_{k \in \{1,\ldots,\xi\}} \sum_{v_j \in V_{M_k}} \left(\alpha * \delta_{bio}(v_j) + (1-\alpha) * \delta_{topo}(v_j)\right) \tag{1}$$

where ξ is the number of PPI networks, V_{M_k} is the set of proteins in M_k, α is a coefficient to trade off biological and topological scores, δ_{bio} and δ_{topo} are the biological and topological scores respectively. The biological score of a protein consists of: (1) the number of links with the subnetwork M_0, (2) the number of links with the subnetwork M_h, and (3) the number of threads among these three subnetworks which contain the current protein. The topological score of a vertex consists of (1) the degree of current vertex; (2) the size of the maximal component that includes the current vertex. As the same with biological score, we adopt a transform techniques by multiplying a coefficient.

References

1. Hu, A.L., Chan, K.C.: Utilizing both topological and attribute information for protein complex identification in ppi networks. IEEE/ACM Trans. Comput. Biol. Bioinform. **10**(3), 780–792 (2013)
2. Li, M., Chen, X., Ni, P., Wang, J., Pan, Y.: Identifying essential proteins by purifying protein interaction networks. In: International Symposium on Bioinformatics Research and Applications (ISBRA), pp. 106–116 (2016)
3. Song, B., Gao, J., Ke, W., Hu, X. Achieving high k-coverage and k-consistency in global alignment of multiple PPI networks. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 303–307 (2016)
4. Malod-Dognin, N., Przulj, N. L-GRAAL: Lagrangian graphlet-based network aligner. Bioinformatics **31**(13), 2182–2189 (2015)
5. Zhao, B., Wang, J., Li, M., Wu, F.X., Pan, Y.: Detecting protein complexes based on uncertain graph model. IEEE/ACM Trans. Comput. Biol. Bioinform. **11**(3), 486–497 (2014)

Constructing an Integrative MicroRNA eQTL Network on Ovarian Cancer: A Label Propagation Approach Utilizing Multiple Networks

Benika Hall, Andrew Quitadamo, and Xinghua Shi

University of North Carolina at Charlotte, Charlotte 28213, USA
{bjohn157,aquitada,x.shi}@uncc.edu

Abstract. Expression quantitative trait loci (eQTL) network construction has been an important task in understanding functional relationships in genomics. In this paper, we construct an integrative microRNA eQTL network based on a label propagation framework using TCGA ovarian cancer data. Label propagation is a robust semi-supervised learning algorithm capable of handling multiple heterogeneous networks reflecting different types of genetic interactions. Elucidation of the interactions involved in multiple networks provide more insight in the dynamics of cancer progression.

Keywords: microRNAs eQTLs · Regulatory networks · Protein protein interaction networks · Network expansion · Label propagation · Ovarian cancer

1 Introduction

Ovarian cancer is the fifth most deadliest cancer among cancer deaths and is responsible for over five percent of cancer deaths in women [1]. MicroRNAs (miRNAs) are small non-coding RNAs that are approximately 22 nucleotides in length and contribute the progression of ovarian cancer through various functional roles such as cell differentiation, apoptosis and tumoriogenesis. Here, we propose a robust semi-supervised learning approach to model the complex relationships between miRNAs, eQTLs and their regulated genes. Expression quantitative trait loci (eQTLs) are genomic regions that can influence gene expression locally or in a distant manner. Thus, we conduct miRNA eQTL analysis to assess the effect of miRNAs on gene expression [2–5].

2 Methods

We downloaded miRNA and gene expression data from TCGA [6], InWeb network [7], a gene regulatory network from RegNetwork database [8] consisting of experimentally verified targets. We conducted eQTL analysis between miRNAs and gene expression and discovering correlations between miRNAs as well as correlations

between genes Lastly, we use our eQTL genes as seed nodes and expand our network with two additional networks, the Inweb and RegNetwork using a label propagation framework.

3 Results

We generated a multi-layered eQTL network including miRNA eQTLs, miRNA correlations, gene correlations, Protein-protein interactions and a gene regulatory network. This integrative network allowed us to capture many facets of gene regulation in ovarian cancer. In the integrated network we have 174 miRNAs and 2,180 genes. These miRNAs and genes are connected through 803 regulatory edges, 1313 protein-protein edges, 9 correlated miRNAs, 18 correlated gene edges and a total of 855 miRNA eQTL edges.

4 Conclusion

We created an integrated miRNA eQTL network utilizing multiple networks. Our integrated network included a miRNA eQTL network, a protein-protein interaction network (InWeb), a gene regulatory network(RegNetwork), and correlation networks on miRNAs and genes respectively. A single miRNA or target usually does not impact the phenotypic outcome individually. To exploit the large scope of regulation, we applied a network based learning approach to integrate multiple networks containing multiple regulatory elements in ovarian cancer.

References

1. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2017. CA Cancer J. Clin. **67**(1), 7–30 (2017)
2. Hall, B., Quitadamo, A., Shi, X., Identifying microrna and gene expression networks using graph communities. Tsinghua Sci. Technol. **21**(2), 176–195 (2016)
3. Quitadamo, A., Tian, L., Hall, B., Shi, X.: An integrated network of microrna and gene expression in ovarian cancer. BMC Bioinform. **16**(5), 1 (2015)
4. Huan, et al.: Genome-wide identification of microrna expression quantitative trait loci. Nat. Commun. **6** (2015)
5. Gamazon, et al.: Genetic architecture of microrna expression: implications for the transcriptome and complex traits. Am. J. Hum. Genet. **90**(6), 1046–1063 (2012)
6. Cancer Genome Atlas Research Network: Integrated genomic analyses of ovarian carcinoma. Nature **474**, 609–615 (2011)

7. Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., et al.: A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. **25**(3), 309–316 (2007)
8. Liu, Z.-P., Wu, C., Miao, H., Wu, H.: Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database 2015:bav095 (2015)

Clustering scRNA-Seq Data Using TF-IDF

Marmar Moussa and Ion Măndoiu

Computer Science and Engineering Department,
University of Connecticut, Storrs, CT, USA
{marmar.moussa,ion}@engr.uconn.edu

Abstract. Single cell RNA sequencing (scRNA-Seq) is critical for understanding cellular heterogeneity and identification of novel cell types. We present novel computational approaches for clustering scRNA-seq data based on the TF-IDF transformation.

Introduction

In this abstract, we propose several computational approaches for clustering scRNA-Seq data based on the Term Frequency - Inverse Document Frequency (TF-IDF) transformation that has been successfully used in the field of text analysis. Empirical evaluation on simulated cell mixtures with different levels of complexity suggests that the TF-IDF methods consistently outperform existing scRNA-Seq clustering methods.

Methods

We compared eight scRNA-Seq methods, including three existing methods and five proposed methods based on the TF-IDF transformation. All methods take as input the raw *Unique Molecular Identifier (UMI)* counts generated using 10X Genomics' Cell-Ranger pipeline [4]. *Existing scRNA-Seq clustering methods* are: the recommended workflow for the Seurat package [3], the Expectation-Maximization (EM) algorithm implemented in the mclust package [2], and a K-means clustering approach similar to that implemented in the CellRanger pipeline distributed by 10X Genomics [1]. Two types of *TF-IDF based methods* were explored. In first type of methods, TF-IDF scores were used to select a subset of the most informative genes that were then clustered with EM and spherical K-means. In the second type all genes were used for clustering, but the expression data was first binarized using a TF-IDF based cutoff. The binary expression level signatures were clustered using: hierarchical clustering with Jaccard distance, and hierarchical clustering with cosine distance with or without an additional cluster aggregation step.

Experimental Setup and Results

To assess accuracy we used mixtures of real scRNA-Seq profiles generated from FACS sorted cells [4]. We selected five cell types: CD8+ cytotoxic T cells (abbreviated as C),

CD4+/CD45RO+ memory T cells (M), CD4+/CD25+ regulatory T cells (R), CD4+ helper T cells (H), and CD19+ B cells (B). We generated mixtures comprised of 5,000 cells sampled from all five cell types in equal proportions. Box-plots of classification accuracy achieved by the eight compared methods are shown in Fig. 1. TF-IDF based hierarchical clustering with cosine distance and cluster aggregation performs better than all other methods, with a mean accuracy of 0.7418, followed by the TF-IDF based spherical K-means, with a mean accuracy of 0.7125.

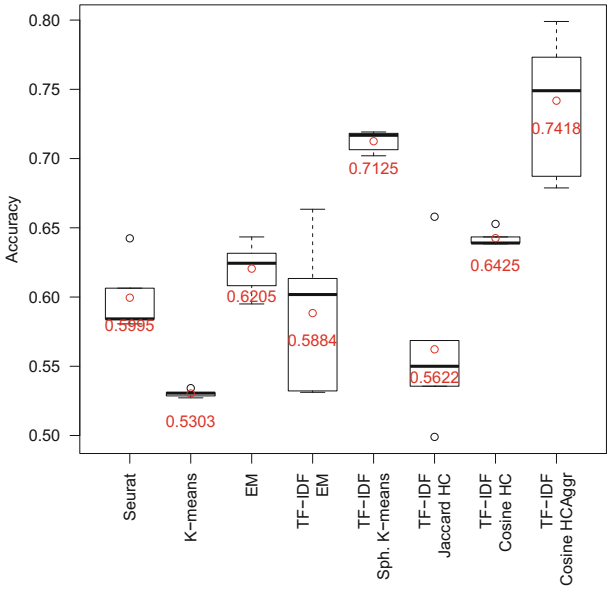

Fig. 1. Accuracy for the B:R:H:M:C datasets with 1:1:1:1:1 ratio.

Acknowledgements. This work was partially supported by NSF Award 1564936 and a UConn Academic Vision Program Grant.

References

1. Cell Ranger R Kit Tutorial. http://s3-us-west-2.amazonaws.com/10x.files/code/cellrangerrkit-PBMC-vignette-knitr-1.1.0.pdf
2. Fraley, C., Raftery, A., Murphy, T., Scrucca, L.: mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. University of Washington, Seattle (2012)
3. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A.: Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. **33**(5), 495–502 (2015)
4. Zheng, G.X.Y., et al.: Massively parallel digital transcriptional profiling of single cells. Nat. Commun. **8**, 14049 (2017)

CircMarker: A Fast and Accurate Algorithm for Circular RNA Detection

Xin Li, Chong Chu, Jingwen Pei, Ion Măndoiu, and Yufeng Wu

Computer Science and Engineering Department,
University of Connecticut, Storrs, CT, USA
{xin.li,chong.chu,jingwen.pei,ion.mandoiu,
yufeng.wug}@uconn.edu

Circular RNA (or circRNA) is a type of RNA which forms a covalently closed continuous loop. It is now believed that circRNA plays important biological roles in some diseases. Within the past several years, several experimental methods, such as RNase R, have been developed to enrich circRNA while degrading linear RNA. Some useful software tools for circRNA detection have been developed as well. However, these tools may miss many circRNA. Also, existing tools are slow for large data because those tools often depend on reads mapping.

In this paper, we present a new computational approach, named CircMarker, based on k-mers rather than reads mapping for circular RNA detection as shown in Fig. 1. The algorithm has two parts, including reference genome proprocessing and annotations (part 1) and circular RNA detection (part 2).

In part 1, CircMarker creates a table for storing the k-mers within the reference genome that are near the exon boundaries as specified by the annotations. The k-mer table is designed to be space-efficient. We only record five types of information for each k-mer, including chromosome index, gene index, transcript index, exon index and part tag. The "part tag" specifies whether a k-mer comes from the head part or the tail part of the exon.

Part 2 is divided into five steps. (1) Sequence reads processing: examine k-mers contained in a read and search for a match in the k-mer table. (2) Filtering by hit number: short exons should be fully covered by the reads more than one time. Otherwise, the reads should be within both boundaries of the hit exons. (3) Filtering by part tags: we collect part tags from start to end, and condense the tags which belong to the same exons based on the number of hits. (4) Calling circRNA: both self-circular case (single exon) and regular-circular case (multiple exon) are considered. In the regular-circular case, we consider if the exon index increases/decreases monotonically and identify the circular joint junction at the position of the first deceasing/increasing position. (5) Refining circular RNA candidates (optional): only the candidates with support number smaller than a predefined threshold will be viewed as correct one.

We use both simulated and real data for evaluation. We compared CircMarker with three other tools, including CIRI [1], Find circ [3], and CIRCexplorer [4] in terms of the number of called circular RNA, accuracy, consensus-based sensitivity, bias and running time. The results are shown in Fig. 1.

- **Simulated Data.** The simulated data is generated by the simulation script released by CIRI. The reference genome is chromosome 1 in human genome (GRCh37). The annotation file is version 18 (Ensembl 73). Two different cases are simulated, including 10X circRNA & 100X linear RNA, and 50X for both circular and linear RNA.
- **Real data: RNase R treated reads with public database.** We choose CircBase [2] as the standard circRNA database of homo sapiens. The reference genome and annotation file come from homo sapiens GRCm37 version 75. The RNA-Seq reads are from SRR901967.
- **Real Data: RNase R treated/untreated Reads.** The reference genome and annotation file are from *Mus Musculus GRCm38 Release79*. RNase R treated/untreated reads are from SRR2219951 and SRR2185851 respectively.

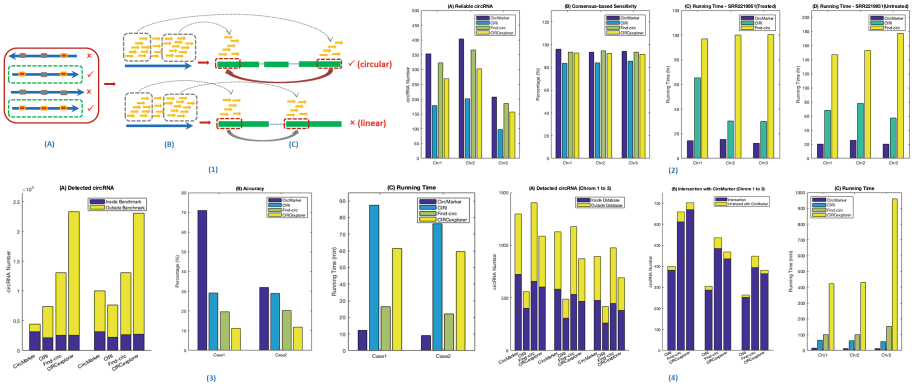

Fig. 1. High Level Approach and Results: (1) High level approach: a fast check for finding circRNA relevant reads, scanning k-mer sequentially from the beginning to the end for each read, and calling circRNA using various criteria and filters. (2) Results of real data based on RNase R treated/untreated reads. (3) Results of simulated data. (4) Results of real data based on RNase R treated reads with public database.

The results show that CircMarker runs much faster and can find more circular RNA than other tools. In addition, CircMarker has higher consensus-based sensitivity and high accuracy/reliable ratio compared with others. Moreover, the circRNAs called by CircMarker often contain most circRNAs called by other tools in the real data we tested. This implies that CircMarker has low bias. CircMarker can be downloaded at: https://github.com/lxwgcool/CircMarker.

References

1. Gao, Y., Wang, J., Zhao, F.: Ciri: an efficient and unbiased algorithm for de novo circular rna identification. Genome Biol. **16**(1), 4 (2015)

2. Glažar, P., Papavasileiou, P., Rajewsky, N.: circbase: a database for circular rnas. RNA **20**(11), 1666–1670 (2014)
3. Memczak, S., Jens, M., Elefsinioti, A., et al.: Circular rnas are a large class of animal rnas with regulatory potency. Nature **495**(7441), 333–338 (2013)
4. Zhang, X.O., Wang, H.B., Zhang, Y., Lu, X., Chen, L.L., Yang, L.: Complementary sequence-mediated exon circularization. Cell **159**(1), 134–147 (2014)

Multiple Model Species Selection for Transcriptomic and Functional Analysis

Kuan-Hung Li[1], Cin-Han Yang[1,2], Tun-Wen Pai[1], Chi-Hua Hu[3], Han-Jia Lin[3], Wen-Der Wang[4], and Yet-Ren Chen[5]

[1] Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan
twp@mail.ntou.edu.tw
[2] Center of Excellence for the Oceans, National Taiwan Ocean University, Keelung, Taiwan
[3] Department of Bioscience and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan
[4] Department of Bioagricultural Science, National Chiayi University, Chiayi City, Taiwan
[5] Agricultural Biotechnology Research Center, Academia Sinica, Taipei, Taiwan

Abstract. Transcriptomic sequencing (RNA-seq) related applications allow for rapid explorations due to their high-throughput and relatively fast experimental capabilities, providing unprecedented progress in gene functional annotation, gene regulation analysis, and environmental factor verification. However, with increasing amounts of sequenced reads and reference model species, the selection of appropriate reference species for gene annotation has become a new challenge. In this study, we proposed a combinatorial approach for finding the most effective reference model species through taxonomic associations and ultra-conserved orthologous (UCO) gene comparisons among species. An online system of multiple species selection (MSS) for RNA-seq differential expression analysis was developed and evaluated. In the designed system, a set of 291 reference model eukaryotic species with comprehensive genomic annotations were selected from the RefSeq, KEGG, and UniProt databases. Using the proposed MSS pipeline, gene ontology and biological pathway enrichment analysis can be efficiently and effectively achieved, especially in the case of transcriptomic analysis of non-model organisms. Regarding the experimental results of selecting appropriate reference model species by analyzing taxonomic relationships and comparing UCOs, accurate evolutionary distances are calculated using sequence alignment and applied to compensate for indistinguishable characteristics of the taxonomic tree. Here, we performed RNA-seq experiments in four non-model species, and the results confirmed that evolutionary distances between species could be ascertained using UCO gene sets. We also performed enrichment analysis of the identified differentially expressed genes using Gene Ontology (GO) and KEGG biological pathway approaches. For example, though GO analysis of *Corbicula fluminea* under hypoxic conditions, we identified additional significant GO terms, including the Notch signaling pathway, cytoskeletal protein binding, and hydrolase activity. These additionally identified GO terms have been found to be associated with hypoxia in previous reported studies. For KEGG biological pathway analysis, additional significant

biological pathways could be also identified, such as the CAM pathway, by increasing the number of appropriate reference species. Therefore, pertinent selection of multiple reference species for transcriptomic analysis can reduce required computational hours and unnecessary searches against the non-redundant gene dataset. In addition, selecting multiple appropriate species as reference model species helps to reduce missing crucial annotation information, allowing for more comprehensive results than those obtained with a single model reference species.

Keywords: RNA-seq · Reference model species · Differential expression analysis · Ultra-conserved orthologous genes · Gene ontology · Biological pathway

RNA Transcriptome Mapping with GraphMap

Krešimir Križanović[1], Ivan Sović[2], Ivan Krpelnik[1], and Mile Šikić[1,3]

[1] Faculty of Electrical Engineering and Computing,
University of Zagreb, Zagreb, Croatia
mile.sikic@fer.hr
[2] Ruđer Bošković Institute, Zagreb, Croatia
[3] Bioinformatics Institute, A*STAR, Singapore, Singapore

The advent of Next Generation Sequencing (NGS) methods has popularized sequencing in various fields of research such as medicine, pharmacy, food technology and agriculture. Aside from DNA sequencing, NGS also enabled RNA sequencing using sequencing-by-synthesis approach. While 3rd generation sequencing technologies are rapidly taking over their share of DNA sequencing market, due to the fact that read length is less important for RNA data analysis, RNA sequencing is still predominately done using NGS. However, it seems likely that at least some aspects of RNA analysis would benefit from increased read length.

Of the currently available RNA-seq aligners BBMap [1] claims to support both PacBio and ONT data, while PacBio GitHub pages offer instructions for working with STAR [2] and GMap [3]. Several available DNA aligners, such as BWA-MEM [4] have been proven to work well with PacBio and ONT data, but they do not offer support for mapping RNA reads to a transcriptome.

In this paper we present an updated version of GraphMap [5] that uses given annotations to generate a transcriptome, and then maps RNA reads to the generated transcriptome using a DNA mapping algorithm. Afterwards, the mapping results are translated back into the genome coordinates. Since initial alignments are calculated for the transcriptome, there is no need to consider spliced alignments and alternative gene splicing. In this way, we can leverage the mapping quality of a proven DNA aligner designed for long and erroneous reads without the need for additional computation to determine exon junctions.

We have compared the new version of GraphMap to three RNA aligners claiming support for 3rd generation sequencing data: BBMap, GMap and STAR. All aligners were tested on three synthetic datasets simulated using a PacBio DNA simulator PBSIM [6]. Since PBSIM is a DNA simulator, to simulate RNA reads it was applied to a transcriptome generated from gene annotations. PBSIM model for CLR reads was used for simulations, and parameters were set for PacBio ROI (Reads of Insert). Alignment results were evaluated by comparing them to MAF files containing information on read origins generated by PBSIM as a part of simulation.

This work has been supported in part by Croatian Science Foundation under the project UIP-11-2013-7353 “Algorithms for Genome Sequence Analysis”.

The results displayed in Table 1 show that GraphMap outperforms other aligners by all criteria successfully aligning a read to all exons from its origin (**hit all**) for over 80% of reads and successfully aligning a read to at least one exon of its origin (**hit one**) for over 90% of the reads. It surpasses the results of other aligners by 5–10% on all datasets.

Table 1. Aligner evaluation results. The table shows the percentage of reads for which alignment overlaps all exons from read origin (**hit all**) and the percentage of reads for which alignment overlaps at least one exon from read origin (**hit one**).

Aligner	STAR		BBMap		GMap		Graphmap	
Dataset	Hit all	Hit one	Hit all	Hit one	Hit all	Hit one	Hit all	Hit one
1	46.7%	47.1%	87.0%	88.1%	84.7%	85.7%	93.5%	94.1%
2	32.1%	35.2%	54.4%	78.4%	73.0%	85.4%	82.0%	94.1%
3	33.1%	35.7%	26.8%	61.2%	70.0%	83.8%	85.7%	94.5%

The research presented in this paper demonstrates that the idea to use an appropriate DNA aligner and gene annotations to map RNA reads to a transcriptome and then to transform the mapping results back to genome coordinates is very feasible. Updated GraphMap clearly outperforms other tested splice aware aligners on all datasets. The results suggest that by implementing splice aware mapping logic into a DNA mapper which works well with third generation sequencing data could also work well for de novo RNA spliced mapping.

Keywords: RNA · Transcriptome · Gene annotations · RNA alignment

References

1. Bushnell, B., Egan, R., Copeland, A., Foster, B., Clum, A., Sun, H., et al: BBMap: A Fast, Accurate, Splice-Aware Aligner (2014)
2. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al.: STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**, 15–21 (2013). doi:10.1093/bioinformatics/bts635
3. Wu, T.D., Watanabe, C.K.: GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21**, 1859–1875 (2005). doi:10.1093/bioinformatics/bti310
4. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013)
5. Sović, I., Šikić, M., Wilm, A., Fenlon, S.N., Chen, S., Nagarajan, N.: Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun. **7**, 11307 (2016). doi:10.1038/ncomms11307
6. Ono, Y., Asai, K., Hamada, M.: PBSIM: PacBio reads simulator–toward accurate genome assembly. Bioinformatics **29**, 119–21 (2013). doi:10.1093/bioinformatics/bts649

A Graph-Based Approach for Proteoform Identification and Quantification Using Homogeneous Multiplexed Top-Down Tandem Mass Spectra

Kaiyuan Zhu[1] and Xiaowen Liu[2,3]

[1] School of Informatics and Computing, Indiana University Bloomington, Bloomington, USA
[2] School of Informatics and Computing, Indiana University-Purdue University Indianapolis, Indianapolis, USA
xwliu@iupui.edu
[3] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, USA

Although protein separation techniques have been significantly advanced, it is still a challenging problem to separate proteoforms with similar weights and similar chemical properties, especially those with the same amino acid sequence, but different post-translational modification (PTM) patterns, in top-down mass spectrometry [1]. Tandem mass spectrometry analysis of two or more proteoforms that are not separated by protein separation methods and have similar molecular masses results in a *multiplexed tandem mass (MTM) spectrum*, which is a superimposing of the tandem mass spectra of the proteoforms [4]. There are two types of MTM spectra: *heterogeneous* multiplexed tandem mass (HetMTM) spectra are generated from proteoforms of two or more different proteins; *homogeneous* multiplexed tandem mass (HomMTM) spectra from proteoforms of the same protein with different PTM patterns.

We focus on the study of the identification and quantification of modified proteoforms using HomMTM spectra, in which purified proteins are often analyzed and the target protein is often known. Let P be a unmodified target protein sequence and S a HomMTM spectrum generated from k modified proteoforms of P. Denote $\mathcal{Q}$ as the set of modified proteoforms of P that match the precursor mass of S. The *HomMTM spectral identification problem* is to find k proteoforms in $\mathcal{Q}$ and their relative abundances such that the peaks (their m/z values and intensities) in spectrum S are best explained [1].

We formulate the HomMTM spectral identification problem as the minimum error k-splittable flow (MEkSF) problem on graphs with vertex capacities, in which each path corresponds to a modified proteoform and the flow on the path corresponds to the relative abundance of the proteoform. The goal is to find a k-splittable flow F with a fixed flow value f (F can be decomposed to k or less than k paths) from the source to the sink in a given graph G such that the sum of the errors on the vertices is minimized.

We prove that the MEkSF problem is NP-hard when k is part of the input and propose a polynomial time algorithm for the problem on layered directed graphs when k is a constant. The algorithm consists of two steps: for a given number k, the packing

step determines a set of flow value candidates for k flows, and the routing step finds out the paths for the k flow values that minimize the sum of errors on vertices. When $k = 2$, we prove that the number of flow value candidates is limited by $|V|$, which is the number of vertices in the graph, and propose an efficient dynamic programming algorithm for solving the routing problem. The total time complexity of the algorithm is $O(l^4 h|V|)$, where l is the largest number of vertices in a layer and h is the number of layers in the graph.

We tested the algorithm on a data set of the histone H4 protein with $3,254$ top-down tandem mass spectra. The mass spectra were deconvoluted using MS-Deconv [3]. After searching the deconvoluted spectra against the histone H4 sequence, the proposed method identified 625 spectra with at least 10 matched fragment ions, of which 441 were matched to single proteoforms and 184 matched to proteoform pairs. For each identified proteoform pair, we computed the difference between the number of fragment ions matched to the pair and that matched to the higher abundance proteoform only. Compared with the higher abundance proteoform, the proteoform pair increased the number of matched fragment ions by at least 10 for 39 of the 184 proteoform pairs. In addition, we computed the difference between the sum of peak intensities explained by the pair and that by the higher abundance proteoform only. Proteoform pairs increased explained peak intensities by at least 20% for 26 spectra compared with single proteoforms.

We also compared the proposed method with MS-Align-E [2] on the histone H4 data set. MS-Align-E identified from the data set $1,037$ spectra, of which 184 were matched to a proteoform pair by the proposed method. For 43 of the 184 spectra, the proposed method increased the number of matched fragment ions by at least 10 compared with MS-Align-E.

Acknowledgement. The research was supported by the National Institute of General Medical Sciences, National Institutes of Health (NIH) through Grant R01GM118470.

References

1. DiMaggio Jr., P.A., Young, N.L., Baliban, R.C., Garcia, B.A., Floudas, C.A.: A mixed integer linear optimization framework for the identification and quantification of targeted post-translational modifications of highly modified proteins using multiplexed electron transfer dissociation tandem mass spectrometry. Mol. Cell. Proteomics **8**, 2527–2543 (2009)
2. Liu, X., Hengel, S., Wu, S., Tolić, N., Paša-Tolić, L., Pevzner, P.A.: Identification of ultra-modified proteins using top-down tandem mass spectra. J. Proteome Res. **12**, 5830–5838 (2013)
3. Liu, X., Inbar, Y., Dorrestein, P.C., Wynne, C., Edwards, N., Souda, P., Whitelegge, J.P., Bafna, V., Pevzner, P.A.: Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. Mol. Cell. Proteomics **9**, 2772–2782 (2010)
4. Wang, J., Perez-Santiago, J., Katz, J.E., Mallick, P., Bandeira, N.: Peptide identification from mixture tandem mass spectra. Mol. Cell. Proteomics **9**, 1476–1485 (2010)

A New Estimation of Protein-Level False Discovery Rate

Guanying Wu[1], Xiang Wan[2], and Baohua Xu[1]

[1] The Dental Center of China-Japan Friendship Hospital, Beijing, China
[2] Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Abstract. In shotgun proteomics, the identification of proteins is a two-stage process: peptide identification and protein inference [1]. In peptide identification, experimental MS/MS spectra are searched against a sequence database to obtain a set of peptide-spectrum matches (PSMs) [2–4]. In protein inference, individual PSMs are assembled to infer the identity of proteins present in the sample [5–7]. Evaluating the statistical significance of the protein identification result is critical to the success of proteomics studies. Controlling the false discovery rate (FDR) is the most common method for assuring the overall quality of the set of identifications. However, the problem of accurate assessment of statistical significance of protein identifications remains an open question [8, 9]. Existing FDR estimation methods either rely on specific assumptions or rely on the two-stage calculation process of first estimating the error rates at the peptide-level, and then combining them somehow at the protein-level. We propose to estimate the FDR in a non-parametric way with less assumptions and to avoid the two-stage calculation process.

We propose a new protein-level FDR estimation framework. The framework contains two major components: the Permutation+BH (Benjamini–Hochberg) FDR estimation method and the logistic regression-based null inference method. In Permutation+BH, the null distribution of a sample is generated by searching data against a large number of permuted random protein database and therefore does not rely on specific assumptions. Then, p-values of proteins are calculated from the null distribution and the BH procedure is applied to the p-values to achieve the relationship of the FDR and the number of protein identifications. The Permutation+BH method generates the null distribution by the permutation method, which is inefficient for online identification. The logistic regression model is proposed to infer the null distribution of a new sample based on existing null distributions obtained from the Permutation+BH method. In our experiment based on three public available datasets, our Permutation+BH method achieves consistently better performance than MAYU, which is chosen as the benchmark FDR calculation method for this study. The null distribution inference result shows that the logistic regression model achieves a reasonable result both in the shape of the null distribution and the corresponding FDR estimation result.

References

1. Nesvizhskii, A.I., Vitek, O., Aebersold, R.: Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Meth. **4**, 787–797 (2007)
2. Eng, J.K., McCormack, A.L., Yates III, J.R.: An approach to correlate tandem mass spectra data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. **5**, 976–989 (1994)
3. Perkins, D.N., Pappin, D.J.C., Creasy, D.M., Cottrell, J.S.: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis **20**, 3551–3567 (1999)
4. Craig, R., Beavis, R.C.: TANDEM: matching proteins with tandem mass spectra. Bioinformatics **20**, 1466–1467 (2004)
5. Nesvizhskii, A.I., Keller, A., Kolker, E., Aebersold, R.: A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. **75**, 4646–4658 (2003)
6. Bern, M., Goldberg, D.: Improved ranking functions for protein and modification-site identifications. In: International Conference on Research in Computational Molecular Biology, pp. 444–458 (2007)
7. Li, Y., Arnold, R., Li, Y., Radivojac, P., Sheng, Q., Tang, H.: A Bayesian approach to protein inference problem in shotgun proteomics. In International Conference on Research in Computational Molecular Biology, pp. 167–180 (2008)
8. Spirin, V., Shpunt, A., Seebacher, J., Gentzel, M., Shevchenko, A., Gygi, S., Sunyaev, S.: Assigning spectrum-specific p-values to protein identifications by mass spectrometry. Bioinformatics **27**, 1128–1134 (2011)
9. Omenn, G.S., Blackwell, T.W., Fermin, D., Eng, J., Speicher, D.W., Hanash, S.M.: Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. Nat. Biotechnol. **24**, 333–338 (2006)

A Generalized Approach to Predicting Virus-Host Protein-Protein Interactions

Xiang Zhou, Byungkyu Park, Daesik Choi, and Kyungsook Han

Department of Computer Science and Engineering,
Inha University, Incheon, South Korea
{jusang486,anrgid6893}@gmail.com
{bpark,khan}@inha.ac.kr

Many computational methods have been developed to predict PPIs, but most of them are intended for PPIs within a same species rather than for PPIs across different species. Motivated by the recent increase in data of virus-host PPIs, a few computational methods have been developed to predict virus-host PPIs, but most of them cannot be applied to new viruses or new hosts that have no known PPIs to the methods. A recent SVM model called DeNovo [1] is perhaps the only one that can predict PPIs of new viruses with a shared host. Protein sequence similarity between different types of viruses or hosts is relatively low, so predicting virus-host PPIs for new viruses or hosts is quite challenging.

We obtained all known PPIs between virus and host from four databases, APID, IntAct, Mentha and UniProt, which use same protein identifiers. As of December 2016, there were a total of 12,157 PPIs between 29 hosts and 332 viruses. For negative data, we obtained protein sequences of major hosts (human, non-human animal, plant, and bacteria) from UniProt, and removed those with a sequence similarity higher than 80% to any positive data.

We constructed several datasets to examine the applicability of our prediction method to new viruses and hosts.

1. Training (TR) and test (TS) sets for assessing the applicability to new viruses
   TR1: 10,955 PPIs between human and any virus except H1N1
   TR2: 11,341 PPIs between human and any virus except Ebola virus
   TR3: 11,617 PPIs between any host and any virus except H1N1
   TR4: 12,007 PPIs between any host and any virus except Ebola virus
   TS1: 381 PPIs between human and H1N1 virus
   TS2: 150 PPIs between human and Ebola virus
2. Training (TR) and test (TS) sets for assessing the applicability to new hosts
   TR5: 11,491 PPIs between human and any virus
   TS5.1: 488 PPIs between non-human animal and any virus
   TS5.2: 17 PPIs between plant and any virus
   TS5.3: 143 PPIs between bacteria and any virus

We built a support vector machine (SVM) model using LIBSVM with the radial basis function as a kernel. The SVM model uses several features of protein sequences: the relative frequency of amino acid triplets (RFAT), frequency difference of amino

acid triplets (FDAT), amino acid composition (AC), and transition, distribution and composition of amino acid groups. The first three features (RFAT, FDAT and AC) are improved features developed in our previous study of single host-virus PPIs [2], and the last three features (transition, distribution and composition) were developed by You et al. [3] for PPIs in a single species.

The SVM model was evaluated in several ways: 10-fold cross validation on several datasets with different ratios of positive to negative data instances and independent testing on new viruses and hosts. In the 10-fold cross validation on three datasets of different ratios of positive to negative data (1:1, 1:2 and 1:3), the best performance (sensitivity = 85%, specificity = 96%, accuracy = 86%, PPV = 86%, NPV = 85%, MCC = 0.71, and AUC = 0.93) was observed in the balanced dataset with 1:1 ratio of positive to negative data. As expected, running the SVM model on unbalanced datasets resulted in lower performances than running it on the balanced dataset.

The model was tested on new viruses using 2 independent datasets of PPIs of H1N1 and Ebola virus, which were not used in training the model. Proteins of H1N1 virus have an average sequence similarity of 9.6% to those of other viruses, and proteins of Ebola virus have a sequence similarity of 10.9% to other viruses. Despite such a low sequence similarity of proteins in test datasets to those in training datasets, the model showed a relatively high performance in independent testing (in datasets TR1-TS1, TR2-TS2, TR3-TS1 and TR4-TS2, it showed accuracies of 78%, 78%, 77% and 82%, respectively).

Likewise, we tested the model on new hosts. A model trained with human-virus PPIs (TR5) was tested on PPIs of viruses with non-human, which include non-human animal (TS5.1), plant (TS5.2) and bacteria (TS5.3). The average sequence similarity of human proteins to non-human animal, plant, and bacteria is lower than 10.7%, but the model showed accuracies of 66%, 68% and 67% in test sets of non-human animal, plant, and bacteria, respectively.

In this study, we developed a general method for predicting PPIs between any virus and any host. In independent testing of the model on new viruses and hosts, it showed a high performance comparable to the best performance of other methods for PPIs between a specific virus and its host. This method will be useful in finding potential PPIs of a new virus or host, for which little information is available. The program and data are available at http://bclab.inha.ac.kr/VirusHostPPI.

References

1. Eid, F.E., ElHefnawi, M., Heath, L.S.: DeNovo: virus-host sequence-based protein-protein interaction prediction. Bioinformatics **32**, 1144–1150 (2016)
2. Kim, B., Alguwaizani, S., Zhou, X., Huang, D.-S., Park, B., Han, K.: An improved method for predicting interactions between virus and human proteins. J. Bioinform. Comput. Biol. **15**(1), 1650024 (2016)
3. You, Z.H., Chan, K.C.C., Hu, P.W.: Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. PLoS One **10**(5), e0125811 (2015)

Deep Learning for Classifying Maize Seeds in Double Haploid Induction Process

Balaji Veeramani, John W. Raymond, and Pritam Chanda

Dow AgroSciences LLC, Indianapolis, IN, USA
{bveeramani,jwraymond,pchanda}@dow.com

1 Introduction

In industrial agricultural breeding, double haploid based generation of inbred maize lines has accelerated the time to market of commercial seed varieties [5]. Traditionally, haploid corn seeds are manually discriminated from the diploid seeds using visual indications of the molecular marker system that is selectively expressed in the embryo region of the diploid seeds. In the industrial scale, there have been two notable automation efforts based on the R1-*nj* marker system [2, 4]. However due to the extensive phenotypic variation of the marker expression [1] and heterogeneity arising from image acquisition in the field, developing computer vision methods to classify seed images is challenging, and approaches robust in recovering haploids are lacking.

2 Results and Discussion

Convolutional neural networks (CNN) have been used successfully for traffic sign recognition, face verification and with autonomous driving vehicles [3]. In this work, we investigate, to our knowledge for the first time, the application of a convolutional network to sort maize haploid seeds from diploids using thousands of images of corn seeds (see Fig. 1). We obtained 4731 corn seed RGB images consisting of 952 haploid and 3779 diploid seeds from several different proprietary maize inbred lines. We train our network using the image dataset that was randomly split into 4021 training (809 haploid and 3212 diploid seeds) and 710 test (143 haploids and 567 diploids) images with 20% haploids in both sets. The training images were further divided into 5-folds to assess its performance under random data splits on unseen data.

We demonstrate deep convolutional networks perform significantly better as compared to several other classifiers that use seed texture, color, and shape features (see Table 1). On the test data set, our network achieved the highest classification accuracy (0.968) among all methods used in our experiments. We looked into the

Balaji Veeramani—Authors acknowledge help/feedback by W. Edsall, S. Cryer, B. John, K. Koehler, G. Tragesser, G. Temnykh, E. Frederickson, and P. Setlur.
Balaji Veeramani and Pritam Chanda contributed equally.

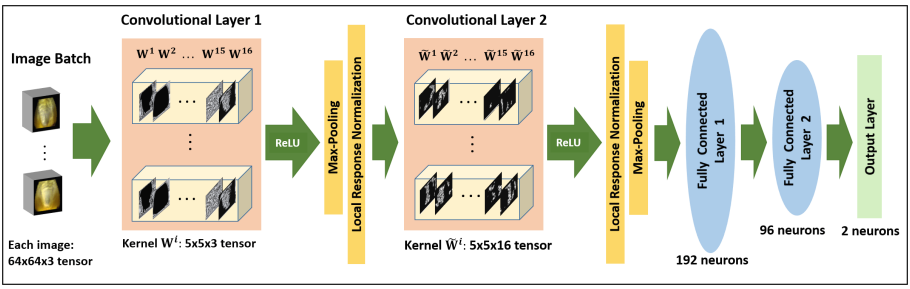

Fig. 1. Convolutional neural network architecture schematic for haploid seed sorting. Input images of the corn seeds are convolved with 16 filter kernels in each convolutional layer, followed by two fully connected layers and an output layer.

misclassifications of our method and the best performing comparative method (SVM) to gain insight into its ability to classify haploid and diploid categories separately in the test dataset. Out of the 567 diploids and 143 haploids in the test dataset, CNN misclassifies 12 haploids as diploids, and 11 diploids as haploids. However, the SVM has a higher tendency to classify haploids as diploids. It classifies 66 haploids as diploids (and 22 diploids as haploids), possibly reflecting dataset class distribution.

Visualizations of the neuronal activations in the convolutional layers indicate the network derives features that are discriminative of embryo regions between haploids and diploids (results not shown here). With the advent of technological advances in agriculture, convolutional networks and other deep learning techniques hold promise for several applications within the agricultural industry.

Table 1. Classification accuracies comparing CNN and other classifiers using texture features (values within brackets indicate results using all features; CV:Cross Validation)

	CNN	SVM	Random forest	Logistic regression
CV	0.961	0.857 (0.836)	0.840 (0.823)	0.749 (0.777)
Train	1.000	0.911 (0.994)	1.000 (0.997)	0.751 (0.786)
Test	0.968	0.876 (0.839)	0.845 (0.824)	0.775 (0.772)

References

1. Kebede, A.Z., Dhillon, B.S., Schipprack, W., Araus, J.L., Bänziger, M., Semagn, K., Alvarado, G., Melchinger, A.E.: Effect of source germplasm and season on the in vivo haploid induction rate in tropical maize. Euphytica **180**(2), 219–226 (2011)
2. Koehler, K.L., Tragesser, G., Swanson, M.: Apparatus and method for sorting plant material. US Patent 9,156,064 (2015)
3. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
4. Modiano, S.H., Deppermann, K.L., Crain, J., Eathington, S.R., Graham, M.: Seed sorter. US Patent 8,189,901 (2012)
5. Prasanna, B., Chaikam, V., Mahuku, G.: Doubled haploid technology in maize breeding: theory and practice. CIMMYT (2012)

PhenoSimWeb: A Web Tool for Measuring and Visualizing Phenotype Similarities Using HPO

Jiajie Peng[1], Hansheng Xue[2], Bolin Chen[1], Qinghua Jiang[3], Xuequn Shang[1], and Yadong Wang[2,4]
[1] School of Computer Science and Technology, Northwestern Polytechnical University, Xi'an, China
[2] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
[3] School of Life Science and Technology, Harbin Institute of Technology, Harbin, China
[4] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

The Human Phenotype Ontology (HPO) was constructed by Robinson *et al.* in 2008, which is one of the most widely used bioinformatics resources [7]. The unified and structured vocabulary of HPO helps to display the phenotypic characteristics, constructs a directly acyclic graph (DAG), and provides a convenient way to study the phenotype similarity.

In recent years, various HPO-based semantic similarity measurements have been proposed to measure the phenotype similarity. Most of these methods are based on the Information Content (IC), including Resnik [6], Schlicker measure [8] and Phenomizer [3]. Besides, PhenomeNet [2] and OWLSim [9] are further developed to calculate two phenotype sets similarity based on simGIC [5]. HPOSim [1] provides an open source package to measure phenotype similarity, which integrates seven widely used HPO-based similarity measurements.

Most of the aforesaid methods are revised based on GO-based similarity measurements, which mainly consider the annotations and topological informations of phenotype terms and neglect the unique features of HPO. Therefore, we proposed a novel method, termed as *PhenoSim*, to calculate the phenotype similarity [4]. Our method consists of denoising model, which model the noises in the patient phenotype data set, and a novel path-constrained Information Content similarity measurement. The whole process of *PhenoSim* can be grouped into three steps: constructing the phenotype network, reducing noise data in patients' phenotype set using PageRank algorithm, and calculating the phenotype set similarities by a novel path-constrained Information Content.

Furthermore, the existing tools of measuring phenotype similarity mainly have two drawbacks: Firstly, existing tools ignores the importance of phenotype text, which are

---

Jiajie Peng, Hansheng Xue — Equal contributor.

often used to describe the symptoms of patients, and none of them allow phenotype text as input. Secondly, none of existing tools supplies interface to visualize the similarity results instead of listing the final similarity value directly. Thus, it is necessary to develop an easy-to-used web application to allow researchers to type in phenotype text and visualize the final phenotype similarity results.

In this paper, we present a novel web tool termed as *PhenoSimWeb*, which is available at 120.77.47.2:8080, to measure HPO-based phenotype similarities and to visualize the result with an easy-to-use graphical interface. Comparing with the existing tools, *PhenoSimWeb* has the following advantages:

- *PhenoSimWeb* offers researchers a novel phenotype semantic similarity measurement which considers the unique features of HPO.
- *PhenoSimWeb* allows researchers to type in the phenotype text that describes phenotype features.
- *PhenoSimWeb* provides an easy-to-use graphical interface to visualize phenotype semantic similarity association.

Acknowledgement. This work was supported the Fundamental Research Funds for the Central Universities (Grant No. 3102016QD003), National Natural Science Foundation of China (Grant No. 61602386 and 61332014), the High-Tech Research and Development Program of China (Grant No. 2015AA020101, 2015AA020108).

References

1. Deng, Y., Gao, L., Wang, B., Guo, X.: Hposim: an r package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. PloS one **10**(2), e0115692 (2015)
2. Hoehndorf, R., Schofield, P.N., Gkoutos, G.V.: Phenomenet: a whole-phenome approach to disease gene discovery. Nucleic Acids Res. **39**(18), e119–e119 (2011)
3. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., Robinson, P.N.: Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am. J. Hum. Genet. **85**(4), 457–464 (2009)
4. Peng, J., Xue, H., Shao, Y., Shang, X., Wang, Y., Chen, J.: Measuring phenotype semantic similarity using human phenotype ontology. In: BIBM, pp. 763–766 (2016)
5. Pesquita, C., Faria, D., Bastos, H., Falcão, A., Couto, F.: Evaluating go-based semantic similarity measures. In: Proceedings of 10th Annual Bio-Ontologies Meeting. vol. 37, p. 38 (2007)
6. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453 (1995)
7. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., Mundlos, S.: The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. **83**(5), 610–615 (2008)
8. Schlicker, A., Domingues, F.S., Rahnenführer, J., Lengauer, T.: A new measure for functional similarity of gene products based on gene ontology. BMC Bioinform. **7**(1), 1 (2006)
9. Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M., Lewis, S.E.: Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol. **7**(11), e1000247 (2009)

An Improved Approach for Reconstructing Consensus Repeats from Short Sequence Reads

Chong Chu, Jingwen Pei, and Yufeng Wu

Computer Science and Engineering Department,
University of Connecticut, Storrs, CT, USA
{chong.chu,jingwen.pei,yufeng.wu}@uconn.edu

A repeat is a segment of DNA that appears multiple times in the genome in an identical or near-identical form. There are many types of repeats such as transposable elements (TEs), tandem repeats, satellite repeats, and simple repeats. Among them, TEs are perhaps the most well-known one. Even though many computational approaches have been developed for constructing consensus repeats, it is still useful to construct repeats directly from reads for complex genomes. Repeats usually have many copies in the genome. For low divergent and high copy number repeats, it is highly likely that k-mers generated from their copies will be identical at the same position. Thus, repeats can be assembled from these high frequent k-mers. RepARK [3] and the original REPdenovo [1] are developed based on this observation. The original REPdenovo outperforms RepARK because it conducts a second-round assembly: it attempts to assemble short contigs in order to form longer consensus repeats based on the reliable prefix-suffix matches of contigs. However, REPdenovo performs less well for highly divergent or low copy number repeats. One reason is that k-mers originated from high divergent regions of a long repeat usually have low frequency, and thus will be filtered out. This leads to fragmented assembled repeats. Another reason is that variations make it difficult to merge the fragmented contigs to form complete repeats. In Fig. 1 (A) and (B), we show two examples to illustrate the situation described above.

In this paper, we propose an improved method (with pipeline shown in Fig. 1(C)) for reconstructing repeat elements from short reads. Similar to the original REPdenovo, our new method also finds and assembles these highly frequent k-mers to form consensus repeat sequences. There are two main improvements in the improved REPdenovo over the original REPdenovo:

- Our new method uses more repeat-related k-mers for repeat assembly, and can assemble longer consensus repeats. Briefly, with high frequent k-mers used as a "reference", low frequent k-mers originated from high divergent regions will be recruited by a "mapping-based alignment" approach.
- Our new method uses a randomized algorithm to generate more accurate consensus k-mers. This improves the quality of the assembled repeats.

Compared to the original REPdenovo and RepARK, our new method can construct more fully assembled repeats in Repbase on both Human and Arabidopsis data, especially for higher divergent, lower copy number and longer repeats. Figure 1(D) shows the comparison between the constructed repeats of the two versions in Repbase

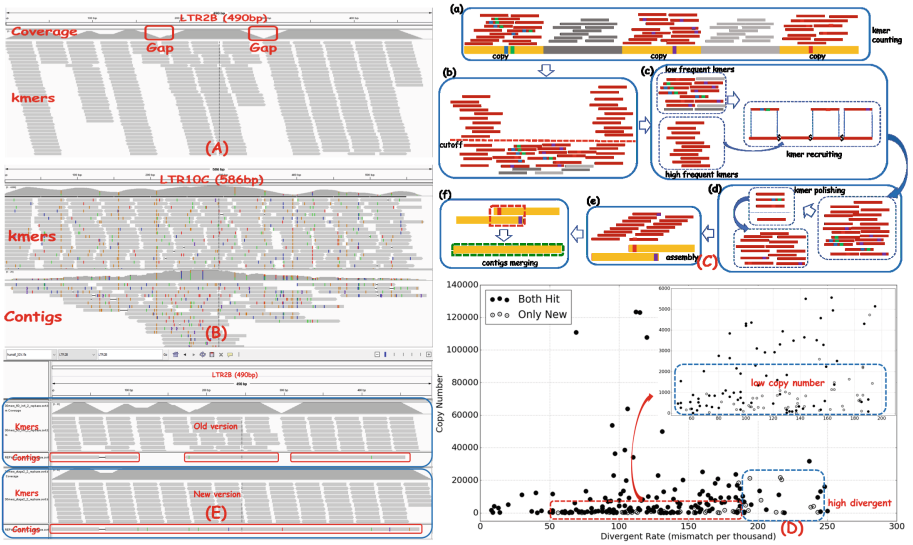

Fig. 1. Observations, pipeline of the method, and results of the improved REPdenovo. (A) Observation one: k-mers from high divergent regions are filtered out and thus form gaps, which leads to fragmented assembled sequences. (B) Observation two: variations make it difficult to assemble long contigs. (C) Pipeline of the improved REPdenovo. (D) Comparison between the original and the improved version of REPdenovo on constructed human repeats in Repbase. (E) One example for comparing the assembly quality on one repeat between the original and the improved REPdenovo.

on Human data. Figure 1(E) illustrates one case that the improved REPdenovo fully construct the repeats while the original REPdenovo fails to. We also apply the new method on Hummingbird data, which has no existing repeat library. Most of the repeats constructed by our new method for Hummingbird can be fully aligned to PacBio long reads. Many of these repeats are long. More than half of the Hummingbird repeats are masked by RepeatMasker, which indicates our assembly works reasonably well. Moreover, many of the assembled repeats are likely to be novel because there are no matches in RepBase. Our new approach has been implemented as part of the REPdenovo software package, which is available for download at https://github.com/Reedwarbler/REPdenovo.

References

1. Chu, C., Nielsen, R., Wu, Y.: Repdenovo: Inferring de novo repeat motifs from short sequence reads. PloS one **11**(3), e0150719 (2016)
2. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J.: Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. **110** (1–4), 462–467 (2005)
3. Koch, P., Platzer, M., Downie, B.R.: RepARK - de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. **42**, e80 (2014)

GRSR: A Tool for Deriving Genome Rearrangement Scenarios from Multiple Unichromosomal Genome Sequences

Dan Wang[1] and Lusheng Wang[1,2]

[1] Department of Computer Science,
City University of Hong Kong, Kowloon, HK
cswangl@cityu.edu.hk
[2] University of Hong Kong Shenzhen Research Institute,
Shenzhen Hi-Tech Industrial Park, Nanshan District,
Shenzhen, People's Republic of China

Sorting genomic permutations by rearrangement operations is a classic problem in studying genome rearrangements. Many tools or algorithms have been proposed for sorting signed genomic permutations [1, 2]. In fact, given a pair of permutations, there are often more than one optimal rearrangement scenarios, especially when the rearrangement distance between this permutation pair is large. And sometimes, for the same pair of permutations, the computed rearrangement scenarios using different tools are not consistent. Hence, how to know whether the calculated scenarios are solid and biologically meaningful becomes an essential task. Up to now, several mechanisms for genome rearrangements have been reported [3, 4]. Statistics analyzes showed that breakpoints are often associated with repetitive elements [5, 6]. There was evidence showing that a reversal can be mediated by a pair of inverted repeats (IRs) [7, 8]. Hence, whether there exist repeats at the breakpoints of rearrangement events may give us a clue on whether the calculated rearrangement scenarios are biologically meaningful.

In this paper, we describe a new tool named GRSR for deriving genome rearrangement scenarios from multiple unichromosomal genome sequences and checking whether there are repeats at the breakpoints of each calculated rearrangement event. The input of the GRSR tool is a set of unichromosomal genome sequences and the output is pairwise rearrangement scenario which is a series of transpositions, block interchanges and reversals. Besides, for each calculated rearrangement event, GRSR checks whether there exist repeats which may mediate this rearrangement event.

The GRSR tool is comprised of four primary steps. Firstly, we use Mugsy [9] to conduct a multiple sequence alignment of the input genomes and the alignment result is in an MAF file. Secondly, as transpositions, block interchanges and reversals happen on sequences which are shared by genomes, we extract the coordinates of core blocks (shared by all of the input genomes) from the MAF file. Thirdly, we utilize the coordinates of core blocks to construct synteny blocks using GRIMM [2] and each input genome will be represented by a signed permutation describing the synteny block order on its chromosome. Lastly, we implement a novel method to compute the pairwise rearrangement scenario which is a series of rearrangement events involved in transforming one genome's permutation into another. The computed rearrangement

scenarios will only include rearrangement events which happen on a single chromosome, such as transpositions, block interchanges and reversals. Given a pair of signed permutations *s* and *d*, the GRSR tool calculate rearrangement scenario from *s* to *d* by merging blocks which are on the same order on *s* and *d*, then detecting and removing obvious (independent) transpositions and block interchanges and finally sorting permutations *s* and *d* by reversals using GRIMM. Once getting a rearrangement event, the GRSR tool will check whether there are repeats at the breakpoints of this event using BLAST [10]. The GRSR tool writes the rearrangement scenarios and whether there are repeats at the breakpoints of each rearrangement event into the *report.txt* file.

We applied the GRSR tool on complete genomes of 28 *Mycobacterium tuberculosis* strains, 24 *Shewanella* strains and 2 *Pseudomonas aeruginosa* strains, respectively. From the results generated by the GRSR tool, we observed that many reversal events were flanked by a pair of inverted repeats so that the two ends of the reversal region remain unchanged before and after the reversal event. We also observed that in other rearrangement operations such transpositions and block interchanges, there exist repeats (not necessarily inverted) at the breakpoints, where the ends remained unchanged before and after the rearrangement operations. In the results for *Pseudomonas aeruginosa* strains, we found an example in which the existence of repeats may explain breakpoint reuse. All the above observations suggest that the conservation of ends could possibly be a popular phenomenon in many types of genome rearrangement events.

References

1. Hannenhalli, S., Pevzner, P.A.: Transforming men into mice (polynomial algorithm for genomic distance problem). In: 36th Annual Symposium on Foundations of Computer Science. Proceedings. IEEE, pp. 581–592 (1995)
2. Tesler, G.: Grimm: genome rearrangements web server. Bioinformatics **18**(3), 492–493 (2002)
3. Darmon, E., Leach, D.R.: Bacterial genome instability. Microbiol. Mol. Biol. Rev. **78**(1), 1–39 (2014)
4. Gray, Y.H.: It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet. **16**(10), 461–468 (2000)
5. Longo, M.S., Carone, D.M., Green, E.D., O'Neill, M.J., O'Neill, R.J., et al.: Distinct retroelement classes define evolutionary breakpoints demarcating sites of evolutionary novelty. BMC Genomics **10**(1), 334 (2009)
6. Sankoff, D.: The where and wherefore of evolutionary breakpoints. J. Biol. **8**(7), 1 (2009)
7. Small, K., Iber, J., Warren, S.T.: Inversion mediated by inverted repeats. Nat. Genet. **16** (1997)
8. Rajaraman, A., Tannier, E., Chauve, C.: Fpsac: fast phylogenetic scaffolding of ancient contigs. Bioinformatics **29**(23), 2987–2994 (2013)
9. Angiuoli, S.V., Salzberg, S.L.: Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics **27**(3), 334–342 (2011)
10. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**(3), 403–410 (1990)

Coestimation of Gene Trees and Reconciliations Under a Duplication-Loss-Coalescence Model

Bo Zhang[1] and Yi-Chieh Wu[2]

[1] Department of Mathematics, Harvey Mudd College,
Claremont, CA, USA
bzhang@hmc.edu
[2] Department of Computer Science, Harvey Mudd College,
Claremont, CA, USA
yjw@cs.hmc.edu

Introduction: Phylogenetic tree reconciliation is fundamental to understanding how genes have evolved within and between species. Given a *gene tree* that depicts how a set of genes has diverged from one another and a *species tree* that depicts how a set of species has speciated, the reconciliation problem proposes a nesting of the gene tree within the species tree and postulates evolutionary events to account for any observed incongruence.

However, within eukaryotes, the most popular reconciliation algorithms consider only a restricted set of evolutionary events, typically modeling only duplications and losses [1, 2] or only coalescences [3, 4]. Recently, the DLCoal model was proposed to unify duplications, losses, and coalescences through an intermediate *locus tree* that describes how new loci are created and destroyed [5]. Here, the locus tree evolves within the species tree according to a duplication-loss model, and the gene tree evolves within the locus tree according to a modified multispecies coalescent model. Two algorithms exist for reconciliations under this model: DLCoalRecon [5], which infers the maximum *a posteriori* reconciliation, and DLCpar [6], which infers a most parsimonious reconciliation. However, both methods assume that the gene tree is known and do not account for errors that may occur during gene tree reconstruction.

To address this challenge, we present DLC-Coestimation, a probabilistic inference method that simultaneously reconstructs the gene tree and reconciles it with the species tree. Given as input a sequence alignment, a species tree, and model parameters including the duplication and loss rate, the population size, and the substitution rate, our algorithm relies on a Bayesian framework to jointly optimize the sequence likelihood and the reconciled tree prior. We show how each term in our inference algorithm corresponds to one component of the underlying generative evolutionary process, and we propose an efficient algorithm for optimizing the overall probability through an iterative hill-climbing procedure combined with Monte Carlo integration.

Results: Our experimental evaluation demonstrates that DLC-Coestimation outperforms existing approaches in ortholog, duplication, and loss inference.

Using a simulated clade of 12 flies, we show that independent reconstruction of the gene tree followed by reconciliation substantially degrades inferences compared to using the true gene tree, even when gene trees are reconstructed with popular top-performing methods. Interestingly, while DLC-Coestimation outperforms DLCoalRecon for every simulation setting, it outperforms DLCpar only for data sets with large amounts of ILS. This finding suggests that our algorithm is better able to handle data sets with low phylogenetic signal, a problem that will become increasingly prevalent as we sequence denser clades.

We also assessed DLC-Coestimation performance on a biological data set of 16 fungi. While all reconciliation methods recover a similar percentage of syntenic orthologs, DLC-Coestimation infers substantially fewer duplications and losses than DLCoalRecon and DLCpar, suggesting that our algorithm is better able to remove spurious duplication and loss events that result from ILS. Furthermore, duplications inferred by DLC-Coestimation are more plausible, with a higher percentage of species overlap post-duplication.

Conclusion: This work demonstrates the utility of coestimation methods for inferences under joint phylogenetic and population genomic models. The DLC-Coestimation software is freely available for download at https://www.cs.hmc.edu/~yjw/software/dlc-coestimation.

References

1. Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. **28**(2), 132–163 (1979)
2. Page, R.D.M.: Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. **43**(1), 58–77 (1994)
3. Kingman, J.F.C.: The coalescent. Stoch. Proc. Appl. **13**(3), 235–248 (1982)
4. Pamilo, P., Nei, M.: Relationships between gene trees and species trees. Mol. Biol. Evol. **5**(5), 568–583 (1988)
5. Rasmussen, M.D., Kellis, M.: Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res. 22, 755–765 (2012)
6. Wu, Y.-C., Rasmussen, M.D., Bansal, M.S., Kellis, M.: Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. Genome Res. **24**(3), 475–486 (2014)

Reconstruction of Real and Simulated Phylogenies Based on Quartet Plurality Inference

Eliran Avni and Sagi Snir

Department of Evolutionary Biology, University of Haifa, Haifa 31905, Israel
ssagi@research.haifa.ac.il

One of the most fundamental tasks in biology is deciphering the history of life on Earth. To achieve that goal, an important step in many phylogenomic analyses is the reconstruction of a tree of ancestor-descendant relationships, a gene tree, for each family of orthologous genes in a dataset. Such analyses have revealed widespread discordance between gene trees [6]. Apart from statistical errors, various mechanisms may lead to incongruences between gene histories, such as hybridization events, duplications and losses in gene families, incomplete lineage sorting, and most importantly, horizontal genetic transfers [4, 9, 11].

Horizontal gene transfer (HGT) is the non-vertical transfer of genes between contemporaneous organisms (as opposed to the standard vertical transmission between parent and offspring). HGT, which is largely mediated by viruses (bacteriophages), plasmids, transposons and other mobile elements, is particularly common in prokaryotes and has been recognized to play an important role in microbial adaptation, with implications in the study of infectious diseases [13]. Estimates of the fraction of genes that experienced HGT vary widely, some as high as 99% [3, 6]. These have led some researchers to question the meaningfulness of the Tree of Life concept [1, 5, 8, 14]. However, despite HGT, that turns evolution into a network of relationships, there is ample evidence that an underlying species tree signal can still be distilled and separated from non tree-like events [2, 6, 7, 10].

In [12], Roch and Snir investigated the feasibility of reconstructing the phylogeny of a four-taxa set - a quartet - using a simple plurality inference rule. Assuming that HGT events are consistent with a Poisson process of a constant rate, they proved that this reconstruction is achieved with high probability if the number of HGT events per gene is $O(\frac{n}{\log n})$ (where n is the number of species). This implies that the number of HGT events can be almost proportional to the number of gene tree edges without destroying the overall tree signal.

In this work we develop the study of the *quartet plurality rule*, by extending it into a complete tree reconstruction scheme. We first complement [12] by finding a lower bound for the probability of simultaneous correct inference of a multitude of quartets, as a function of the size of the species set, the number of gene trees, and the frequency of HGT events. Since every phylogeny is uniquely determined by its induced quartets, accurate reconstruction of the entire set of quartets implies accurate phylogenetic reconstruction, that can be done in this case in polynomial time. Next, we show via detailed simulations, that even when the number of HGT events is much larger than

what the theory of [12] dictates, the plurality inference rule still enables accurate tree reconstruction. In the last part of the paper, we demonstrate that the plurality rule can be a viable tool for real data phylogenetic reconstruction, by applying the above theoretical principles to two sets of prokaryotes. The constructed phylogenies of these two sets are shown to be comparable with (and complementary better than) other suggested evolutionary trees in a number of tests.

Based on our analysis, some interesting questions arise. From a theoretical perspective, our ability to reconstruct accurate phylogenies in practice despite surprisingly high rates of HGT, suggest that the known upper bound for HGT rates that still enable successful tree reconstruction can be further improved. In addition, it is noteworthy that weights were also incorporated in the reconstruction scheme used in this paper. Since only three types of weights were tested, it would be desirable to explore new weighting functions that may be beneficial to the accuracy of tree reconstruction.

References

1. Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R.L., Doolittle, W.F.: Do orthologous gene phylogenies really support tree-thinking? BMC Evol. Biol. **5**, 33 (2005)
2. Beiko, R.G., Harlow, T., Ragan, M.: Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. USA **102**, 14332–14337 (2005)
3. Dagan, T., Martin, W.: The tree of one percent. Genome Biol. **7**(10), 118 (2006)
4. Doolittle, W.F.: Phylogenetic classification and the universal tree. Science **284**(5423), 2124–2129 (1999)
5. Doolittle, W.F., Bapteste, E.: Pattern pluralism and the tree of life hypothesis. Proc. Natl. Acad. Sci. USA **104**, 2043–2049 (2007)
6. Galtier, N. and V. Daubin. Dealing with incongruence in phylogenomic analyses. Philos. Trans. R. Soc. Lond. B Biol. Sci. **363**, 4023–4029 (2008)
7. Ge, F., Wang, L., Kim, J.: The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. **3**, e316 (2005)
8. Gogarten, J.P., Townsend, J.P.: Horizontal gene transfer, genome innovation and evolution. Nat. Rev. Micro. **3**(9), 679–687 (2005)
9. Ochman, H., Lawrence, J.G., Groisman, E.A.: Lateral gene transfer and the nature of bacterial innovation. Nature **405**(6784), 299–304 (2000)
10. Koonin, E.V., Puigbó, P., Wolf, Y.I.: Comparison of phylogenetic trees and search for a central trend in the forest of life. J. Comput. Biol. **18**(7), 917–924 (2011)
11. Maddison, W.P.: Gene trees in species trees. System. Biol. **46**(3), 523–536 (1997)
12. Roch, S., Snir, S.: Recovering the tree-like trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. In: RECOMB, pp. 224–238 (2012)
13. Smets, B.F., Barkay, T.: Horizontal gene transfer: perspectives at a crossroads of scientific disciplines. Nat. Rev. Micro. **3**(9), 675–678 (2005)
14. Zhaxybayeva, O., Lapierre, P., Gogarten, J.: Genome mosaicism and organismal lineages. Trends Genet. **20**, 254–260 (2004)

On the Impact of Uncertain Gene Tree Rooting on Duplication-Transfer-Loss Reconciliation

Soumya Kundu[1] and Mukul S. Bansal[1,2]

[1] Department of Computer Science and Engineering,
University of Connecticut, Storrs, USA
soumya.kundu@uconn.edu
[2] Institute for Systems Genomics, University of Connecticut, Storrs, USA
mukul.bansal@uconn.edu

Duplication-Transfer-Loss (DTL) reconciliation is one of the most effective techniques for studying the evolution of gene families and inferring evolutionary events. Given the evolutionary tree for a gene family, i.e., a *gene tree*, and the evolutionary tree for the corresponding species, i.e., a *species tree*, DTL reconciliation compares the gene tree with the species tree and reconciles any differences between the two by proposing gene duplication, horizontal gene transfer, and gene loss events. DTL reconciliations are generally computed using a parsimony framework where each evolutionary event is assigned a cost and the goal is to find a reconciliation with minimum total cost [1–3]. The resulting optimization problem is called the *DTL-reconciliation problem*.

The standard formulation of the DTL-reconciliation problem requires the gene tree and the species tree to be rooted. However, while species trees can generally be confidently rooted (using outgroups, for example), gene trees are often difficult to root. As a result, the gene trees used for DTL reconciliation are often unrooted. When provided with an unrooted gene tree, existing DTL-reconciliation algorithms and software first find a root for the unrooted gene tree that yields the minimum reconciliation cost and then use the resulting rooted gene tree for the reconciliation. However, there is a critical flaw in this approach: Many gene trees have multiple optimal roots, and yet, only a single optimal root is randomly chosen to create the rooted gene tree and perform the reconciliation. Here, we perform the first in-depth analysis of the impact of uncertain gene tree rooting on DTL reconciliation and provide the first computational tools to quantify and negate the impact of gene tree rooting uncertainty.

To properly account for rooting uncertainty, we define a *consensus reconciliation*, which summarizes the different reconciliations across all optimal rootings of an unrooted gene tree and makes it possible to identify those aspects of the reconciliation that are conserved across all optimal rootings. We study basic structural properties of consensus reconciliations and analyze a large biological data set of over 4500 gene families from a broadly sampled set of 100 predominantly prokaryotic species [4]. Our analysis focuses on several fundamental aspects of DTL reconciliation with unrooted gene trees including prevalence of multiple optimal rootings, structure of optimal roots in multiply rooted gene trees, impact of gene tree error and evolutionary event costs, information content of consensus reconciliations, and conservation of event and mapping assignments in consensus reconciliations.

Our experimental results show that a large fraction of gene trees have multiple optimal rootings and that gene tree error significantly increases the fraction of multiply rooted gene trees. The prevalence of multiple optimal rootings is also heavily influenced by gene tree size, with smaller gene trees more likely to have multiple optimal roots. An analysis of the placement of optimal roots shows that multiple roots often, but not always, appear clustered together in the same region of the gene tree. This a highly desirable property since it maximizes the information content, or size, of consensus reconciliations and also makes it easier to estimate the "true" root position. A detailed study of the computed consensus reconciliations reveals that most aspects of the reconciliation, i.e., event and mapping assignments, remain conserved across the multiple rootings, showing that unrooted gene trees can be meaningfully reconciled even after accounting for multiple optimal roots. Our analysis also uncovers several interesting patterns in the reconciliations of singly rooted and multiply rooted gene trees.

The results of our experimental analysis have important implications for the application of DTL reconciliation in evolutionary studies, and the techniques introduced in this work make it possible to systematically avoid incorrect evolutionary inferences caused by incorrect or uncertain gene tree rooting. Our tools for computing consensus reconciliations have been implemented into the phylogenetic reconciliation software package RANGER-DTL, freely available from http://compbio.engr.uconn.edu/software/RANGER-DTL/.

References

1. Tofigh, A., Hallett, M.T., Lagergren, J.: Simultaneous identification of duplications and lateral gene transfers. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**(2), 517–535 (2011)
2. Doyon, J.P., Scornavacca, C., Gorbunov, K.Y., Szöllosi, G.J., Ranwez, V., Berry, V.: An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: Tannier, E. (ed.) RECOMB-CG. LNCS, vol. 6398, pp. 93–108. Springer, Berlin (2010)
3. Bansal, M.S., Alm, E.J., Kellis, M.: Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics **28**(12), 283–291(2012)
4. David, L.A., Alm, E.J.: Rapid evolutionary innovation during an archaean genetic expansion. Nature **469**, 93–96 (2011)

Biomedical Event Extraction via Attention-Based Bidirectional Gated Recurrent Unit Networks Utilizing Distributed Representation

Lishuang Li, Jia Wan, Jieqiong Zheng, and Jian Wang

Dalian University of Technology, School of Computer Science and Technology, Dalian, China
lilishuang314@163.com

The Bacteria Biotope event extraction (BB) task [1] as the one of biomedical event extraction task has been put forward in the BioNLP Shared Task in 2016. The purpose of the BB task is to study the interaction mechanisms of the bacteria with their environment from genetic, phylogenetic and ecology perspectives. The methods based on shallow machine learning methods for BB event extraction need to extract the manual features. However, the construction of complex hand-designed features mainly relies on preferred experience and knowledge. Furthermore, manual efforts may hurt the generalization performance of the system and lead to over-design. Deep learning methods provide an effective way to reduce the number of handcrafted features. But the approaches take all words as equally important and are not able to capture the most important semantic information in a sentence.

In this paper, we propose a novel Bidirectional Gated Recurrent Unit (BGRU) Networks framework based on attention mechanism, using the corpus from the BioNLP'16 Shared Task on BB task. The BGRU networks as a deep learning framework can reduce the number of handcrafted features and the attention mechanism can take advantage of the important information in the sentence. Simultaneously, we employ a biomedical domain-specific word representation training model, which merges relevant biomedical information including stem, chunk, entity and part-of-speech (POS) tags into word embeddings. The system architecture for event extraction based on attention-based BGRU can be summarized in Fig. 1. Firstly, the Shortest Path enclosed Tree (SPT) between two entities is obtained by GENIA Dependency parser (GDEP) [2] and the SPT is extended to the dynamic extended tree (DET) [3], which can accurately encode the input information. Secondly, the DET is mapped to embeddings which are concatenated by the word embeddings, POS embeddings and distance embeddings. Thirdly, a recurrent neural network with attention-based BGRU is established to acquire the hidden layer. Then, the significant information in a sentence is obtained by a weight vector, which could learn word features automatically. Therefore, a sentence feature can be gained by multiplying the weight vector. Lastly, we utilize a softmax function to predict the label for classification.

The experimental results on the BioNLP-ST'16 BB-event corpus show that our attention mechanism and word representation conditioned BGRU can achieve an

F-score of 57.42%. Without using the complex hand-designed features, our system outperforms the previous state-of-the-art BB-event system.

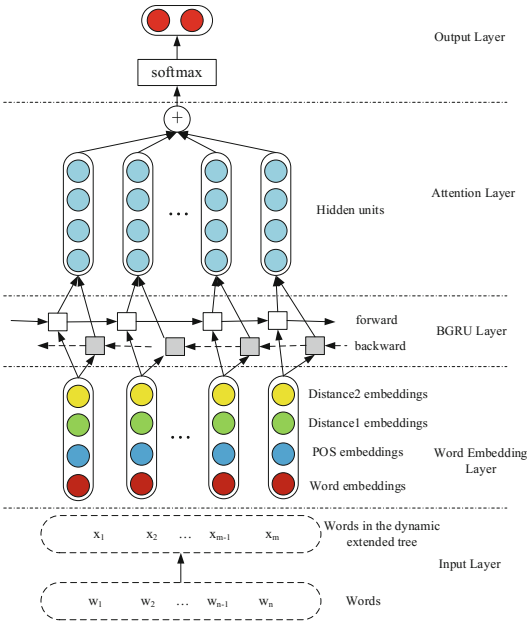

Fig. 1. The architecture of attention-based BGRU

Acknowledgments. The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China under No. 61672126, 61173101.

References

1. Deleger, L., Bossy, R., Chaix, E., Ba, M., Ferré, A., Bessieres, P., Nédellec, C.: Overview of the bacteria biotope task at BioNLP shared task. In: Proceedings of the 4th BioNLP Shared Task Workshop, pp. 12–22. The Association for Computational Linguistics, Berlin (2016)
2. Sagae, K., Tsujii, J.I.: Dependency parsing and domain adaptation with LR models and parser ensembles. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1044–1050. Prague (2007)
3. Li, L., Jin, L., Zheng, J., Zhang, P., Huang, D.: The protein-protein Interaction extraction based on full texts. In: 2014 IEEE International Conference on Bioinformatics and Biomedicine, pp. 497–500. IEEE, Belfast (2014)

Efficient Computation of Motif Discovery on Intel Many Integrated Core (MIC) Architecture

Kaiwen Huang[1], Zhiqiang Zhang[1], Runxin Guo[1], Xiaoyu Zhang[1], Shunyun Yang[1], Xiangke Liao[1], Yutong Lu[2], Quan Zou[3], and Shaoliang Peng[1]

[1] School of Computer Science, National University of Defense Technology, Changsha 410073, China
pengshaoliang@nudt.edu.cn
[2] National Supercomputer Center in Guangzhou, Guangzhou 510275, China
[3] School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
zouquan@nclab.net

Identifying meaningful patterns (*i.e.*, motifs) from biological sequences is an important problem and a major challenge in bioinformatics research. A motif [1] is a nucleotide or amino-acid sequence pattern that recurs in different DNA or protein sequences and has a biological significance. In recent years, it has emerged a large number of computational algorithms for motif discovery which can be categorized into two groups, including word-based (string-based) methods and probabilistic methods [1]. Word-based methods mostly exhaustive enumerate in their computation and probabilistic methods employ probabilistic sequence models where the model parameters are optimized by maximum-likelihood principle or Bayesian inference. Probabilistic methods have the advantage of few parameters and are more appropriate for finding longer or more general motifs especially for prokaryotes, whose motifs are generally longer than eukaryotes.

MEME (Multiple EM for Motif Elicitation) [2] is one of the currently widely-used algorithms based on maximum-likelihood principle for *de novo* motif discovery [3]. The algorithm consists of two stages: starting point searching and EM. The time complexity of MEME is $O(N^2 \times L^2)$, where N is the number of input sequences and L is the average length of each sequence. However, the high computational cost constrains MEME for handling large datasets [4]. To accelerate motif discovery algorithm, most of previous approaches focus on using parallelization on distributed workstations, Graphics Processing Unit (GPU) and Field Programmable Gate Arrays (FPGA). Farouk *et al*. parallelized the Brute Force algorithm targeted on FPGAs [5]. Marchand *et al*. scaled Dragon Motif Finder (DMF) to IBM Blue Gene/P using mixed-mode MPI-OpenMP programming [6]. mCUDA–MEME is a parallel implementation of MEME running on multiple GPUs using CUDA programming model [7].

Kaiwen Huang, Zhiqiang Zhang, Runxin Guo, Xiaoyu Zhang, Shunyun Yang—These authors contributed equally to this work.

Intel Many Integrated Core (MIC) Architecture [8] is the latest co-processor computer architecture developed by Intel, which combines many Intel processor cores onto a single chip to support the most demanding high-performance computing applications. It is a brand-new many-core architecture that delivers massive thread parallelism, data parallelism, vectorization, and memory bandwidth in a CPU form factor for high throughput workloads.

In this paper, we accelerate MEME algorithm targeted on Intel Many Integrated Core (MIC) Architecture to harness the powerful compute capability of MIC and present a parallel implementation of MEME called MIC-MEME base on hybrid CPU/MIC computing framework. Since the starting point searching stage is the runtime bottleneck of the sequential MEME algorithm, our method focuses on parallelizing the starting point searching method and improving iteration updating strategy of the algorithm. And in EM stage, the M step and E step of EM algorithm are simply parallelized using OpenMP. We also take advantage of the 512 bit vectorization unit to get good performance out of the Intel MIC Architecture.

To evaluate the performance of MIC-MEME, the real datasets with different numbers of sequences and base pairs (bps) were used. MIC-MEME produces the same results as sequential MEME. And it has achieved significant speedups of 26.6 for ZOOPS model and 30.2 for OOPS model on average for the overall runtime when benchmarked on the experimental platform with two Xeon Phi 3120 coprocessors. Furthermore, MIC-MEME shows good scalability with respect to dataset size and the number of MICs. And MIC-MEME has been compared favorably with mCUDA-MEME and BoBro2.0. As the result shows, MIC-MEME is average 2.2 times faster than mCUDA-MEME and MIC-MEME absolutely outperforms BoBro2.0. Comparing with the other methods, we can improve the efficiency of MEME algorithm without losing accuracy and our method which makes full use of computing resources is faster and robustness. With the increase of biological data, we hope the efficient motif discovery of MIC-MEME will be able to help the bioresearch work. Source code can be accessed at https://github.com/hkwkevin28/MIC-MEME.

Acknowledgments. This work was supported by NSFC Grants U1435222, 61625202, 61272056, National Key R&D Program 2016YFC1302500, 2016YFB0200400, and Guangdong Provincial Department of Science and Technology under grant No. 2016B090918122.

References

1. Das, M.K., Dai, H.K.: A survey of DNA motif finding algorithms. BMC Bioinform. **8**, Suppl. **7**(7), S21 (2007)
2. Bailey, T.L., Elkan, C., Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. **2**, 28–36 (1994)
3. Bailey, T.L., et al., MEME: Discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. **34**(Web Server issue), 369–373 (2006)

4. Hu, J., Li, B., Kihara, D.: Limitations and potentials of current motif discovery algorithms. Nucleic Acids Res. **33**(33), 4899–4913 (2005)
5. Farouk, Y., Eldeeb, T., Faheem, H.: Massively Parallelized DNA Motif Search on FPGA. InTech. (2011)
6. Marchand, et al.: Highly scalable ab initio genomic motif identification (2011)
7. Liu, Y., Schmidt, B., Maskell, D.L.: An Ultrafast scalable many-core motif discovery algorithm for multiple gpus. In: IEEE International Symposium on Parallel and Distributed Processing Workshops & Phd Forum, pp. 428–434 (2011)
8. Jeffers, J., Reinders, J.: Intel Xeon Phi coprocessor high-performance programming. Morgan Kaufmann Publishers Inc. pp. xvii–xviii (2013)

Predicting Diabetic Retinopathy and Identifying Interpretable Biomedical Features Using Machine Learning Algorithms

Hsin-Yi Tsao[1,2], Pei-Ying Chan[3,4], and Emily Chia-Yu Su[1]

[1] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan
{g658101006,emilysu}@tmu.edu.tw
[2] Division of Endocrinology and Metabolism, Department of Internal Medicine, Sijhih Cathay General Hospital, New Taipei City, Taiwan
[3] Department of Occupational Therapy and Healthy Aging Center, Chang Gung University, Taoyuan, Taiwan
chanp@mail.cgu.edu.tw
[4] Department of Psychiatry, Linkou Chang Gung Memorial Hospital, Taoyuan, Taiwan

Abstract. Diabetic retinopathy (DR) was found to be a frequent comorbid complication to diabetes. The risk factors of DR were investigated extensively in the past studies, but it remains unknown which risk factors were more associated with the DR than others. If we can detect the DR related risk factors more accurately, we can then exercise early prevention strategies for diabetic retinopathy in the most high-risk population. Thus, using computational approaches to predict diabetes mellitus becomes crucial to support medical decision making.

The purpose of this study is to build a prediction model for the DR in type 2 diabetes mellitus using data mining techniques. First, data consisting of 106 DR and 430 normal patients were collected from the "Diabetes Mellitus Shared Care" database in a private hospital in northern Taiwan. We randomly selected 160 patients were from normal group to combine with DR group, and formed a balanced data set. Ten variables, including systolic blood pressure (SBP), diastolic blood pressure (DPB), body mass index (BMI), age, gender, duration of diabetes, family history of diabetes, self-monitoring blood glucose (SMBG), exercise, and insulin treatment, were extracted. Four machine learning algorithms including support vector machines (SVM), decision trees, artificial neural networks, and logistic regressions, were used to predict diabetic retinopathy.

Among these variables, insulin treatment, SBP, DPB, BMI, age, and duration of diabetes showed significant differences between DR and normal groups. Experimental results demonstrated SVM achieved the best prediction performance with 0.839, 0.795, 0.933, and 0.724 in area under curve, accuracy, sensitivity, and specificity, respectively. The aim of this study is not only to achieve an accurate prediction performance, but also to generate an interpretable model for clinical practice. Table 1 and Fig. 1 demonstrated the interpretable

rules generated by logistic regression and decision tree, respectively. Use of insulin and longer duration of DM were major predictors of DR in the decision tree models. If duration of DM increases by 1 year, the odds ratio to have DMR is increased by 9.3%. The odds ratio to have DR is increased by 3.561 times for patients who use insulin compared to patients who do not use insulin. In summary, our method identifies use of insulin and duration of diabetes as novel interpretable features to assist with clinical decisions in identifying the high-risk populations for diabetic retinopathy.

Keywords: Diabetic mellitus retinopathy · Machine learning · Decision support

Table 1. Odds ratio estimates of duration and insulin variables.

Odds ratio estimates		
Effect		Point estimate
Duration		1.093
Insulin	Y vs. N	3.561

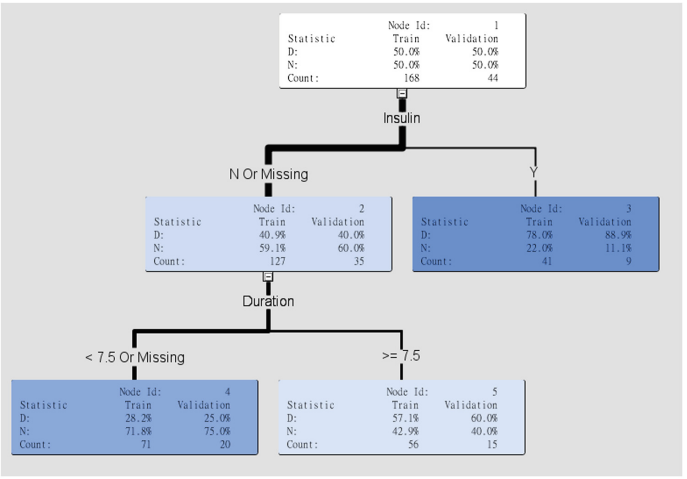

Fig. 1. Interpretable rules for clinical practice generated by decision tress.

Big Data Analysis for Evaluating Bioinvasion Risk

Chenyu Wang[1], Shengling Wang[1], and Liran Ma[2]

[1] College of Information Science and Technology,
Beijing Normal University, Beijing, China
henryascend@gmail.com, wangshengling@bnu.edu.cn
[2] Department of Computer Science, Texas Christian University,
Fort Worth, TX, USA,
l.ma@tcu.edu

Abstract. The global maritime trade makes species get translocated through ballast water and biofouling. We propose a biosecurity triggering mechanism to evaluate the bioinvasion risk of ports. To that aim, we take advantage of big data to compute the invaded risk and construct a species invasion network (SIN). The former is used to evaluate the incoming bioinvasion risk while the latter is employed to estimate the invasion risk spreading capability of a port through *s*-core decomposition.

1 Introduction

Nowadays, people's daily lives are heavily dependent on global maritime trade. However, marine invasive species and viruses would cause side effects in terms of environment and human health, which lead to huge losses of lives and economy [3].

To address the issue of aquatic bioinvasion, one mainstream countermeasure is to propose suggestions for biomarker identification [1, 2] and bioinvasion management. However, the existing biosecurity suggestions only considered the invaded risk of a port and neglected its role of being a *stepping-stone*.

In this paper, we propose a biosecurity triggering mechanism to address the issues of the existing work. In our biosecurity triggering mechanism, once the bioinvasion risk of a port is larger than a given threshold, biosecurity controls should be triggered. To that aim, we take advantage of the automatic identification system (AIS) data, the ballast water data, and the marine ecoregion data to compute the invasion risk between any two ports, based on which the invaded risk is calculated and a species invasion network (SIN) is constructed. Through *s*-core decomposition of SIN, the ports whose *s*-core are higher are identified as the ones transmit bioinvasion risks to others more easily. We found two regions, namely the Western Europe and the Asia-Pacific, which are estimated to be bioinvasion risk intensive regions through our big data analysis.

2 Basis for Our Analysis

For any port j, its invaded risk (i.e. $P_j(Inv)$) is the accumulating invasion risks over all shipping routes passing through it [5], i.e.

$$P_j(Inv) = 1 - \Pi_i[1 - P_{ij}(Inv)] \tag{1}$$

where $P_{ij}(Inv)$ denotes the invasion risk from port i to j.

A SIN can be depicted by a directed graph, namely $S = (V, E, W)$, consisting of a set V of nodes (i.e., ports), a set E of edges (i.e., shipping routes) and the weight $w_{ij} \in W$ ($w_{ij} = P_{ij}(Inv)$ of edge $e_{ij} \in E$ denoting the invasion risk from ports i to j.

According to the description above, both the invaded risk and SIN involve $P_{ij}(Inv)$ $(i,j \in V)$. In this paper, we use the model proposed in [5] to calculate $P_{ij}(Inv)$.

To figure out the potential of a port to spread invaded species to others, we need to dig out the transmission power of each node in SIN, which is closely related to the topological property of each port in SIN. We think k-core decomposition is an efficient tool to analyze the structure of complex networks. Larger values of the index k correspond to nodes with larger degree and more central position. According to the algorithm in [4], we can deduce the s-cores of SIN. Seattle, Tokyo and Lima are the top 3 ports ranked by their value of s-shell.

3 Biosecurity Triggering Method

The main idea of the proposed biosecurity triggering method is to trigger bioinvasion treatment according to the bioinvasion risk of each port. As we introduced above, the bioinvasion risk is estimated in light of both the invaded risk of port and its ability of further spreading invaded species. The former is the incoming risk while the latter is the outgoing one. Therefore, we can trigger the corresponding bioinvasion control on a port j based on the following simple criterion:

$$R(j) = \alpha\widetilde{P}_j(Inv) + (1 - \alpha)\widetilde{s}(j) \geq T \tag{2}$$

where $R(j)$ is the bioinvasion risk of port j, and $\widetilde{P}_j(Inv)$ and $\widetilde{s}(j)$ are respectively the normalized $P_j(Inv)$ (the invaded risk of port j calculated using (1)) and the normalized s-shell value of that port; $0 \leq \alpha \leq 1$ is the tradeoff weight. Smaller α means more attention should be paid on the stepping-stone invasion and otherwise, the invaded risk should be obtained more concern. T is the given threshold to help judging whether a bioinvasion treatment should be triggered.

We found two regions, namely the Western Europe and the Asia-Pacific, are bioinvasion risk intensive regions. The result is consistent with the real-world data. Hence, our analysis basically accords with the real-world marine bioinvasion status.

References

1. Cai, Z., Goebel, R., Salavatipour, M.R., Lin, G.: Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. BMC Bioinform. **8**(1), 206 (2007)
2. Cai, Z., Heydari, M., Lin, G.: Iterated local least squares microarray missing value imputation. J. Bioinform. Comput. Biol. **4**(05), 935–957 (2006)
3. Cai, Z., Zhang, T., Wan, X.F.: A computational framework for influenza antigenic cartography. PLoS Comput. Biol. **6**(10), e1000949 (2010)
4. Eidsaa, M., Almaas, E.: S-core network decomposition: A generalization of k-core analysis to weighted networks. Physical Rev. E **88**(6), 062819 (2013)
5. Seebens, H., Gastner, M.T., Blasius, B.: The risk of marine bioinvasion caused by global shipping. Ecology Lett. **16**(6), 782–790 (2013)

Drug Response Prediction Model Using a Component Based Structural Equation Modeling Method

Sungtae Kim[1] and Taesung Park[2]

[1] Interdisciplinary Program in Bioinformatics,
Seoul National University, Seoul, Korea
[2] Department of Statistics, Seoul National University San 56-1,
Sillim-dong, Gwanak-gu, Seoul 151-742, Korea
tspark@stats.snu.ac.kr

The liver is made up of many different types of cells. Mutations in those cells can be developed into several different forms of tumors known as cancers. For this reason, it is hard to expect a single type of liver cancer treatment to have a favorable prognosis for all cancer patients. If we can diagnose and classify the patients who are expected to have good responses to a single therapeutic drug, it will help to reduce the time on choosing an appropriate therapeutic drug for each patient. Therefore, building a decent prediction model became important for an effective treatment. Up to date, several methods such as linear/logistic regression (LR), support vector machine (SVM), random forest (RF) have been used for building prediction models [1–3]. However, occasionally, these methods oversight the biological pathway information with relations between metabolites, proteins, or DNAs.

In this paper, we propose building of prediction model using component based structured equation modeling method which uses the peptide to protein biological structure. Our peptide level data were generated by Multiple Reaction Monitoring (MRM) mass spectrometry for liver cancer patients. MRM is a highly sensitive and selective method for targeted quantitation of peptide abundances in complex biological samples. The advantage of component based structured equation modeling is that it can generate latent variables. These latent variables are not observable but can be inferred from other observed variables. Using latent variables, we can collapse unstructured data into structured data. These latent variables provide more feasible explanation on the results. In our case, multiple peptides can be merged into a protein which is represented as a latent variable. Our proposed schematic model using component based structural equation modeling for MRM data is shown in Fig. 1.

We applied the component based structural equation model to MRM data of liver cancer patients. In our MRM data, there are 124 proteins induced by 231 peptides MRM data. Each protein contains at least one peptides. We identified candidate proteins for a drug Sorafenib response for liver cancer patients. The selected candidate proteins included APOC4, CD163, CD5L, JCHAIN, SERPING1, and RBP4. These proteins were reported as possible cancer biomarkers [4, 5]. Also, CD5L was well known as a liver cancer biomarker [6, 7]. Using these proteins, we evaluated our proposed Sorafeib prediction model by the area under the curve (AUC) score. Also, we

compared the performance of our model with generalized linear models with and without ridge penalty. The performance of our model showed a slightly higher AUC score 0.96 compared to 0.949 AUC score of the generalized linear model with ridge penalty.

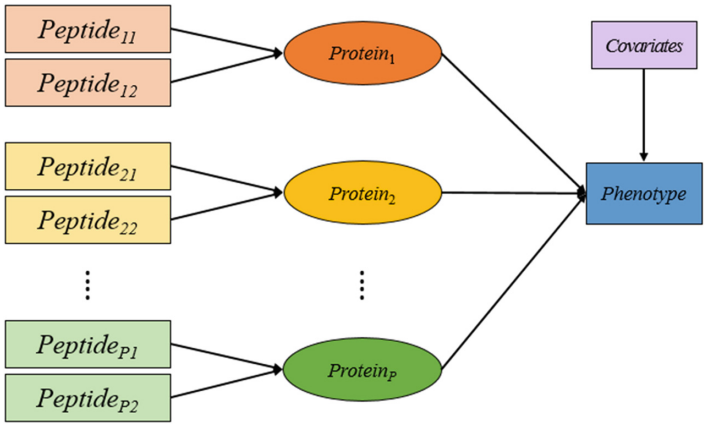

Fig. 1. Proposed component based structural equation model using MRM data.

References

1. Visser, H., le Cessie, S., Vos, K., Breedveld, F.C., Hazes, J.M.: How to diagnose rheumatoid arthritis early: a prediction model for persistent (erosive) arthritis. Arthritis Rheum. **46**(2), 357–365 (2002)
2. Spitz, M.R., Etzel, C.J., Dong, Q., Amos, C.I., Wei, Q., Wu, X., Hong, W.K.: An expanded risk prediction model for lung cancer. Cancer Prev. Res **1**(4), 250–254 (2008)
3. Huang, C.L., Liao, H.C., Chen, M.C.: Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Syst. Appl. **34**(1), 578–587 (2008)
4. Gray, J., Chattopadhyay, D., Beale, G.S., Patman, G.L., Miele, L., King, B.P., Reeves, H.L.: A proteomic strategy to identify novel serum biomarkers for liver cirrhosis and hepatocellular cancer in individuals with fatty liver disease. BMC Cancer **9**(1), 271 (2009)
5. Braconi, C., Meng, F., Swenson, E., Khrapenko, L., Huang, N., Patel, T.: Candidate therapeutic agents for hepatocellular cancer can be identified from phenotype associated gene expression signatures. Cancer **115**(16), 3738–3748 (2009)
6. Chambers, A.G., Percy, A.J., Simon, R., Borchers, C.H.: MRM for the verification of cancer biomarker proteins: recent applications to human plasma and serum. Expert Rev. Proteomics **11**(2), 137–148 (2014)
7. Rabouhans, J.: A radiologist's guide to the modified Response Evaluation Criteria in Solid Tumours (mRECIST) assessment of therapy for hepatocellular carcinoma. European Congress of Radiology (2011)

Contents

Prediction of Time to Tumor Recurrence in Ovarian Cancer: Comparison of Three Sparse Regression Methods 1
Mahsa Lotfi, Burook Misganaw, and Mathukumalli Vidyasagar

Histopathological Diagnosis for Viable and Non-viable Tumor Prediction for Osteosarcoma Using Convolutional Neural Network 12
Rashika Mishra, Ovidiu Daescu, Patrick Leavey, Dinesh Rakheja, and Anita Sengupta

Relating Diseases Based on Disease Module Theory 24
Peng Ni, Min Li, Ping Zhong, Guihua Duan, Jianxin Wang, Yaohang Li, and FangXiang Wu

Reconstructing One-Articulated Networks with Distance Matrices 34
Kuang-Yu Chang, Yun Cui, Siu-Ming Yiu, and Wing-Kai Hon

Computational Methods for the Prediction of Drug-Target Interactions from Drug Fingerprints and Protein Sequences by Stacked Auto-Encoder Deep Neural Network 46
Lei Wang, Zhu-Hong You, Xing Chen, Shi-Xiong Xia, Feng Liu, Xin Yan, and Yong Zhou

Analysis of Paired miRNA-mRNA Microarray Expression Data Using a Stepwise Multiple Linear Regression Model 59
Yiqian Zhou, Rehman Qureshi, and Ahmet Sacan

IsoTree: De Novo Transcriptome Assembly from RNA-Seq Reads (Extended Abstract) 71
Jin Zhao, Haodi Feng, Daming Zhu, Chi Zhang, and Ying Xu

Unfolding the Protein Surface for Pattern Matching 84
Heng Yang, Chunyu Zhao, and Ahmet Sacan

Estimation of Rates of Reactions Triggered by Electron Transfer in Top-Down Mass Spectrometry 96
Michał Aleksander Ciach, Mateusz Krzysztof Łącki, Błażej Miasojedow, Frederik Lermyte, Dirk Valkenborg, Frank Sobott, and Anna Gambin

Construction of Protein Backbone Fragments Libraries on Large Protein Sets Using a Randomized Spectral Clustering Algorithm 108
Wessam Elhefnawy, Min Li, Jianxin Wang, and Yaohang Li

Mapping Paratope and Epitope Residues of Antibody Pembrolizumab via Molecular Dynamics Simulation 120
Wenping Liu and Guangjian Liu

A New 2-Approximation Algorithm for rSPR Distance 128
Zhi-Zhong Chen, Youta Harada, and Lusheng Wang

An SIMD Algorithm for Wraparound Tandem Alignment 140
Joshua Loving, John P. Scaduto, and Gary Benson

PhAT-QTL: A Phase-Aware Test for QTL Detection. 150
Meena Subramaniam, Noah Zaitlen, and Jimmie Ye

Unbiased Taxonomic Annotation of Metagenomic Samples 162
Bruno Fosso, Graziano Pesole, Francesc Rosselló, and Gabriel Valiente

Genetic Algorithm Based Beta-Barrel Detection for Medium Resolution Cryo-EM Density Maps. 174
Albert Ng and Dong Si

Mining K-mers of Various Lengths in Biological Sequences. 186
Jingsong Zhang, Jianmei Guo, Xiaoqing Yu, Xiangtian Yu, Weifeng Guo, Tao Zeng, and Luonan Chen

Coestimation of Gene Trees and Reconciliations Under a Duplication-Loss-Coalescence Model 196
Bo Zhang and Yi-Chieh Wu

A Median Solver and Phylogenetic Inference Based on DCJ Sorting 211
Ruofan Xia, Jun Zhou, Lingxi Zhou, Bing Feng, and Jijun Tang

Addressing the Threats of Inference Attacks on Traits and Genotypes from Individual Genomic Data 223
Zaobo He, Yingshu Li, Ji Li, Jiguo Yu, Hong Gao, and Jinbao Wang

Phylogenetic Tree Reconciliation: Mean Values for Fixed Gene Trees 234
Paweł Górecki, Alexey Markin, Agnieszka Mykowiecka, Jarosław Paszek, and Oliver Eulenstein

Computer Assisted Segmentation Tool: A Machine Learning Based Image Segmenting Tool for TrakEM2 246
Augustus N. Tropea, Janey L. Valerio, Michael J. Camerino, Josh Hix, Emmalee Pecor, Peter G. Fuerst, and S. Seth Long

Accelerating Electron Tomography Reconstruction Algorithm ICON Using the Intel Xeon Phi Coprocessor on Tianhe-2 Supercomputer 258
Zihao Wang, Yu Chen, Jingrong Zhang, Lun Li, Xiaohua Wan, Zhiyong Liu, Fei Sun, and Fa Zhang

Modeling the Molecular Distance Geometry Problem Using Dihedral Angles. 270
Michael Souza, Carlile Lavor, and Rafael Alves

What's Hot and What's Not? - Exploring Trends in Bioinformatics Literature Using Topic Modeling and Keyword Analysis 279
Alexander Hahn, Somya D. Mohanty, and Prashanti Manda

Structure Modeling and Molecular Docking Studies of Schizophrenia Candidate Genes, Synapsins 2 (*SYN2*) and Trace Amino Acid Receptor (*TAAR6*) 291
Naureen Aslam Khattak, Sheikh Arslan Sehgal, Yongsheng Bai, and Youping Deng

Accurate Prediction of Haplotype Inference Errors by Feature Extraction 302
Rogério S. Rosa and Katia S. Guimarães

Detecting Change Points in fMRI Data via Bayesian Inference and Genetic Algorithm Model. 314
Xiuchun Xiao, Bing Liu, Jing Zhang, Xueli Xiao, and Yi Pan

Extracting Depression Symptoms from Social Networks and Web Blogs via Text Mining 325
Long Ma, Zhibo Wang, and Yanqing Zhang

A Probabilistic Approach to Multiple-Instance Learning. 331
Silu Zhang, Yixin Chen, and Dawn Wilkins

Net2Image: A Network Representation Method for Identifying Cancer-Related Genes 337
Bolin Chen, Yuqiong Jin, and Xuequn Shang

Computational Prediction of Influenza Neuraminidase Inhibitors Using Machine Learning Algorithms and Recursive Feature Elimination Method . . . 344
Li Zhang, Haixin Ai, Qi Zhao, Junfeng Zhu, Wen Chen, Xuewei Wu, Liangchao Huang, Zimo Yin, Jian Zhao, and Hongsheng Liu

Differential Privacy Preserving Genomic Data Releasing via Factor Graph . . . 350
Zaobo He, Yingshu Li, and Jinbao Wang

In Silico Simulation of Signal Cascades in Biomedical Networks Based on the Production Rule System. 356
Sangwoo Kim and Hojung Nam

Using the Precision Medicine Analytical Method to Investigate the Impact of the Aerobic Exercise on the Hypertension for the Middle-Aged Women . . . 362
Wei Zhou, Guangdi Liu, Jun Luo, Tingran Zhang, and Le Zhang

Understanding Protein-Protein Interface Formation Mechanism in a New Probability Way at Amino Acid Level 368
Yongxiao Yang and Xinqi Gong

Detecting Potential Adverse Drug Reactions Using Association Rules and Embedding Models 373
Kai Guo, Hongfei Lin, Bo Xu, Zhihao Yang, Jian Wang, Yuanyuan Sun, and Kan Xu

Genome-Wide Analysis of Response Regulator Genes in *Solanum lycopersicum* 379
Jun Cui, Ning Jiang, Jun Meng, and Yushi Luan

A Fully Automatic Geometric Parameters Determining Method for Electron Tomography 385
Yu Chen, Zihao Wang, Lun Li, Xiaohua Wan, Fei Sun, and Fa Zhang

Evaluating the Impact of Encoding Schemes on Deep Auto-Encoders for DNA Annotation 390
Ning Yu, Zeng Yu, Feng Gu, and Yi Pan

Metabolic Analysis of Metatranscriptomic Data from Planktonic Communities 396
Igor Mandric, Sergey Knyazev, Cory Padilla, Frank Stewart, Ion I. Măndoiu, and Alex Zelikovsky

NemoLib: A Java Library for Efficient Network Motif Detection 403
Andrew Andersen and Wooyoung Kim

A Genetic Algorithm for Finding Discriminative Functional Motifs in Long Non-coding RNAs 408
Brian L. Gudenas and Liangjiang Wang

Heterogeneous Cancer Cell Line Data Fusion for Identifying Novel Response Determinants in Precision Medicine 414
Wojciech Czaja and Jeremiah Emidih

Agent-Based in Silico Evolution of HCV Quasispecies 420
Alexander Artyomenko, Pelin B. Icer, Pavel Skums, Sumathi Ramachandran, Yury Khudyakov, and Alex Zelikovsky

Modeling the Spread of HIV and HCV Infections Based on Identification and Characterization of High-Risk Communities Using Social Media 425
Deeptanshu Jha, Pavel Skums, Alex Zelikovsky, Yury Khudyakov, and Rahul Singh

Author Index 431