PNImodeler: Web Server for Inferring Protein Binding Nucleotides from Sequence Data

Jinyong Im, Narankhuu Tuvshinjargal, Byungkyu Park,
Wook Lee, and Kyungsook Han*

Department of Computer Science and Engineering
Inha University, Incheon, South Korea
khan@inha.ac.kr

Interactions between DNA and proteins are essential to many biological processes such as transcriptional regulation and DNA replication. With the increased availability of structures of protein-DNA complexes, several computational studies have been conducted to predict DNA binding sites in proteins. However, little attempt has been made to predict protein binding sites in DNA. From an extensive analysis of protein-DNA complexes obtained from the Protein Data Bank (PDB), we identified powerful features of DNA and protein sequences which can be used in predicting protein binding sites in DNA sequences. The features can be classified into three types: original DNA sequence, DNA sequence fragments from the original DNA sequence, and protein sequence interacting with the DNA. The original DNA sequence is represented by its nucleotide composition. DNA sequence fragments are represented by nucleotide triplet composition, normalized position, molecular mass, molecular pKa and interaction propensity of nucleotide triplets. For protein, which is an interaction partner of DNA, the sum of normalized position of twenty amino acids in the sequence and dipeptide composition are represented.

We have developed two SVM models to predict protein binding nucleotides in DNA. One model uses DNA sequence data alone and predicts all potential binding sites with unknown protein partners. The other model uses both DNA and protein sequences to predict protein binding nucleotides with the specific protein. One SVM model that used DNA sequence data alone achieved a sensitivity of 73.4%, a specificity of 64.8%, an accuracy of 68.9% and a Matthews correlation coefficient (MCC) of 0.382 with a test dataset that was not used in training. Another SVM model that used both DNA and protein sequences achieved a sensitivity of 67.6%, a specificity of 74.3%, an accuracy of 71.4% and MCC of 0.418. The SVM model that used both DNA and protein sequences yielded better overall performance than the model that used DNA sequence alone.

The SVM models have been implemented as a web server called PNImodeler (Protein-Nucleic acid Interaction modeler), and the web server is available at http://bclab.inha.ac.kr/pnimodeler. PNImodeler will be useful to find protein-binding sites in DNA with unknown structure. To the best of our knowledge, this is the first attempt to predict protein-binding DNA nucleotides with sequence data alone.

* Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 367, 2014.

A MCI Decision Support System Based on Ontology

Xiaowei Zhang[1,*], Yang Zhou[1], Bin Hu[2,1], Jing Chen[1], and Xu Ma[1]

[1] School of Information Science and Engineering, Lanzhou University, Lanzhou, China
[2] College of Electronic Information and Control Engineering,
Beijing University of Technology, Beijing, China
{zhangxw,yzhou11,bh,max2012,jchen10}@lzu.edu.cn

Mid Cognitive Impairment (MCI) [1] threatens the health of the elderly around the world and could progress to Alzheimer's Disease (AD) [2] with a high risk. It is necessary to detect MCI earlier to reduce the occurrence rate of AD.

So we develop a Decision Support System (DSS) for detecting subjects with MCI. This system is based on fMRI-Bayesian ontology (FB-Ontology) combined with Bayesian networks algorithm. The DSS employs Functional Magnetic Resonance Imaging (fMRI) techniques to distinguish MCI patients from normal controls (NC). We preprocess fMRI data and calculate path length, global efficiency and hub node features based on the automated anatomical labeling (AAL) template for DSS. By using Bayesian networks algorithm, our DSS could provide uncertain reasoning results for clinicians. Meanwhile, the FB-Ontology acts like a bridge between reasoning engine and low-level database in system, it could provide transparent, unified, normalized, shareable knowledge to users. Finally, we select 22 subjects with MCI and 18 normal controls from Alzheimer's Disease Neuroimaging Initiative (ADNI) [3]. Using a 5-fold cross validation method for training and testing, the system could reach an average classification rate of 90%.

References

1. Gauthier, S., et al.: Mild cognitive impairment. The Lancet 367(9518), 1262–1270 (2006)
2. Petersen, R.C.: Mild Cognitive Impairment: Aging to Alzheimer's Disease. Brain 127(1), 231–233 (2004)
3. http://www.loni.ucla.edu/adni/

* Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 368, 2014.

Context Similarity Based Feature Selection Methods for Protein Interaction Article Classification*

Yifei Chen[1], Yuxing Sun[1], and Ping Hou[2]

[1] School of Information Science, Nanjing Audit University, 86 Yushan Rd(W), Nanjing, P.R. China
[2] Fondazione Bruno Kessler (FBK-irst), Trento, Italy

An overwhelming amount of biological articles are published daily online as a result of growing interest in biological research, especially the study of protein-protein interactions. It is essential to classify which articles describe the protein interactions. Therefore study on automatic protein interaction articles classification (IAC) has become a task with practical significance to the text classification in biological domain.

Since the feature space in text classification is high dimensional, feature selection techniques are widely used for reducing the dimensionality of features to speed up the computation of the classifier. However, the existing feature selection methods are mostly based on the term frequency or document frequency. These approaches are context independent, that is, they do not utilize the context information in a document when judging the importance of features, such as word order, multi-word phrases and semantic relationships, which are important for the IAC tasks. Hence, based on the study of four well-known frequency based feature selection methods, Gini Index (GI), Document Frequency (DF), Class Discriminating Measure (CDM) and Accuracy Balanced (Acc2), we propose four context similarity based feature selection methods, GI_{cs}, DF_{cs}, CDM_{cs} and $Acc2_{cs}$, to introduce the similarity measure of context multi-word phrases.

In order to evaluate the performances of the proposed context similarity based feature selection methods, two data sets ($Data_{BCII}$ and $Data_{BCIII}$) are used in our experiments, which are both extracted from the BioCreAtIvE challenges. The experimental results reveal that all the context similarity based methods outperform the corresponding frequency based methods in terms of the micro-F1 measure. On the $Data_{BCII}$, when top 4.3% features are selected, GI_{cs} acquires the highest $F1$ measure value, which effectively improves the performance when all the features are used by 2.54. And on the $Data_{BCIII}$ when the top 7.4% are selected, CDM_{cs} acquires the highest $F1$ measure value, which improves the performance when all the features are used by 1.12. Moreover, through the analysis on the comparison of selected features and the dimension reduction rate, the proposed methods provide better performances by bring more distinguishing information with the fewer selected features for the text classifier.

* This work is supported by the National Natural Science Foundation of China (No.61202135), the Natural Science Foundation of Jiangsu Province (No.BK2012472) and the Qing Lan Project.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 369, 2014.

Genome-Wide Analysis of Transcription Factor Binding Sites and Their Characteristic DNA Structures

Zhiming Dai[1], Dongliang Guo[1], Xianhua Dai[1], and Yuanyan Xiong[2,3]

[1] Department of Electronics and Communication Engineering,
School of Information Science and Technology,
Sun Yat-Sen University, Guangzhou 510006, China
[2] State Key Laboratory for Biocontrol, Sun Yat-Sen University, Guangzhou 510275, China
[3] SYSU-CMU Shunde International Joint Research Institute, Shunde, China
mody0911@gmail.com

Transcription factors (TF) regulate gene expression by binding DNA regulatory regions. Transcription factor binding sites (TFBSs) are conserved not only in primary DNA sequences but also in DNA structures [1,2]. However, the global relationship between TFs and their preferred DNA structures remains to be elucidated. In this paper, we have developed a computational method to generate a genome-wide landscape of TFs and their characteristic binding DNA structures in *Saccharomyces cerevisiae*. TFBSs are conserved in different DNA structures, independent of sequence conservation. We revealed DNA structural features for different TFs. The structural conservation shows positional preference in TFBSs. Structural levels of DNA sequences are correlated with TF-DNA binding affinities. Our findings will have implications in understanding TF regulatory mechanisms.

References

[1] Parker, S.C., Hansen, L., Abaan, H.O., Tullius, T.D., Margulies, E.H.: Local DNA topography correlates with functional noncoding regions of the human genome. Science 324, 389–392 (2009)

[2] Broos, S., Soete, A., Hooghe, B., Moran, R., van Roy, F., De Bleser, P.: PhysBinder: improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties. Nucleic Acids Res., W531–W534 (2013)

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 370, 2014.

A Comparative Study of Disease Genes and Drug Targets in the Human Protein Interactome

Jingchun Sun, Kevin Zhu, W. Jim Zheng, and Hua Xu

School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Most complex diseases are caused by variation in many genes, which are defined as disease genes [1]. Medicines (drugs) are major choices to treat the diseases or reduce their symptoms as they act through interacting with some proteins [2]. These proteins are defined as drug targets. Thus, disease genes contribute to the pathology of one disease while drug targets are critical for the efficacy of disease treatment. However, the interrelationship between the disease genes and drug targets is not clear.

In this study, we collected disease genes from GWAS catalog database and drug targets from DrugBank and TTD databases. We compared them and found that, though their intersections were small, disease genes were significantly enriched in targets compared to their enrichment in the human protein-coding genes. We further compared network properties of the proteins encoded by disease genes and drug targets in the human interactome. The results showed that the drug targets tended to have a higher degree, a higher betweenness, and a lower clustering coefficient. Additionally, we observed a clear fraction increase of disease proteins or drug targets in the near neighborhood compared with the randomized genes, which is consistent with previous results [3].

The study first comprehensively compared the disease genes and drug targets. The results provide network characteristics for designing computational strategies to predict novel drug targets and drug repurposing.

References

[1] Lander, E.S., Schork, N.J.: Genetic dissection of complex traits. Science 265, 2037–2048 (1994)

[2] Schreiber, S.L.: Target-oriented and diversity-oriented organic synthesis in drug discovery. Science 287, 1964–1969 (2000)

[3] Yildirim, M.A., Goh, K.I., Cusick, M.E., Barabasi, A.L., Vidal, M.: Drug-target network. Nat. Biotechnol. 25, 1119–1126 (2007)

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 371, 2014.

Efficient Identification of Endogenous Mammalian Biochemical Structures

Mai A. Hamdalla, Reda A. Ammar, and Sanguthevar Rajasekaran*

Computer Science and Engineering Department, University of Connecticut, Connecticut, USA
rajasek@engr.uconn.edu

Metabolomics is the comprehensive, qualitative, and quantitative study of all the small molecules, called metabolites, in an organism [1]. A major challenge in metabolomics is the interpretation of the vast amount of data produced by the high-throughput techniques used for information extraction and data interpretation [2]. The existence of several on-line chemical structure databases has provided a vital support for molecular identification by allowing the search for candidate compounds using experimentally determined features with computationally simulated features. Such searches often result in a large number of false positives, making identification of the compound under investigation extremely difficult. Hence, cheminformatics methods are needed to efficiently search such large chemical databases and potentially identify unknown endogenous biochemical compounds. Several methods [3–6] have been developed with the objective of discriminating between candidate structures that are synthetic and those that are biochemical. In the talk, we will present an efficient cheminformatics tool that uses known endogenous mammalian biochemicals and graph matching methods to identify endogenous mammalian biochemical structures in chemical structure space.

References

1. Villas-Bôas, S.G., Bruheim, P.: The potential of metabolomics tools in bioremediation studies. Omics: A Journal of Integrative Biology 11, 305–313 (2007)
2. Kertesz, T., Hill, D.W., Albaugh, D., Hall, L., Hall, L., Grant, D.F.: Database searching for structural identification of metabolites in complex biofluids for mass spectrometry-based metabonomics. Bioanalysis 1, 1627–1643 (2009)
3. Nobeli, I., Ponstingl, H., Krissinel, E.B., Thornton, J.M.: A structure-based anatomy of the E.coli metabolome. Journal of Molecular Biology 334, 697–719 (2003)
4. Gupta, S., Aires-de Sousa, J.: Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. Molecular Diversity 11, 23–36 (2007)
5. Peironcely, J.E., Reijmers, T., Coulier, L., Bender, A., Hankemeier, T.: Understanding and classifying metabolite space and metabolite-likeness. PloS One 6 (2011)
6. Hamdalla, M.A., Mandoiu, I.I., Hill, D.W., Rajasekaran, S., Grant, D.F.: BioSM: A chemoinformatics tool for identifying biochemical structures in chemical structure space. Journal of Chemical Information and Modeling (2012)

* Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 372, 2014.

LncRNA2Function: A Comprehensive Resource for Functional Investigation of Human lncRNAs Based on RNA-seq Data

Qinghua Jiang[1,*], Rui Ma[2,*], Jixuan Wang[3], Xiaoliang Wu[3], Shuilin Jin[4], Jiajie Peng[2], Renjie Tan[2], Tianjiao Zhang[2], Yu Li[1], and Yadong Wang[2,**]

[1] School of Life Science and Technology,
Harbin Institute of Technology, Harbin, Heilongjiang 150001, China
[2] School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, Heilongjiang 150001, China
ydwang@hit.edu.cn
[3] School of Software, Harbin Institute of Technology,
Harbin, Heilongjiang 150001, China
[4] Department of Mathematics, Harbin Institute of Technology,
Harbin, Heilongjiang, 150001, China

The GENCODE project has collected over 10,000 human long non-coding RNA (lncRNA) genes. However, the vast majority of them remain to be functionally characterized. Computational investigation of potential functions of human lncRNA genes is helpful to guide further experimental studies on lncRNAs. In this study, based on expression correlation between lncRNAs and protein-coding genes across 19 human normal tissues, we used the hypergeometric test to functionally annotate a single lncRNA or a set of lncRNAs with significantly enriched functional terms among the protein-coding genes that are co-expressed with the lncRNA(s). The functional terms include all nodes in the Gene Ontology (GO) and 4,380 human biological pathways collected from 12 pathway databases. We mapped 9,625 human lncRNA genes to GO terms and biological pathways. Finally, we developed the first ontology-driven tool named lncRNA2Function, which enables researchers to browse the lncRNAs associated with a specific functional term, the functional terms associated with a specific lncRNA, or to assign functional terms to a set of human lncRNA genes such as a cluster of co-expressed lncRNAs. The lncRNA2Function is freely available at http://mlg.hit.edu.cn/lncrna2function.

[*] Contributed equally.
[**] Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 373–374, 2014.

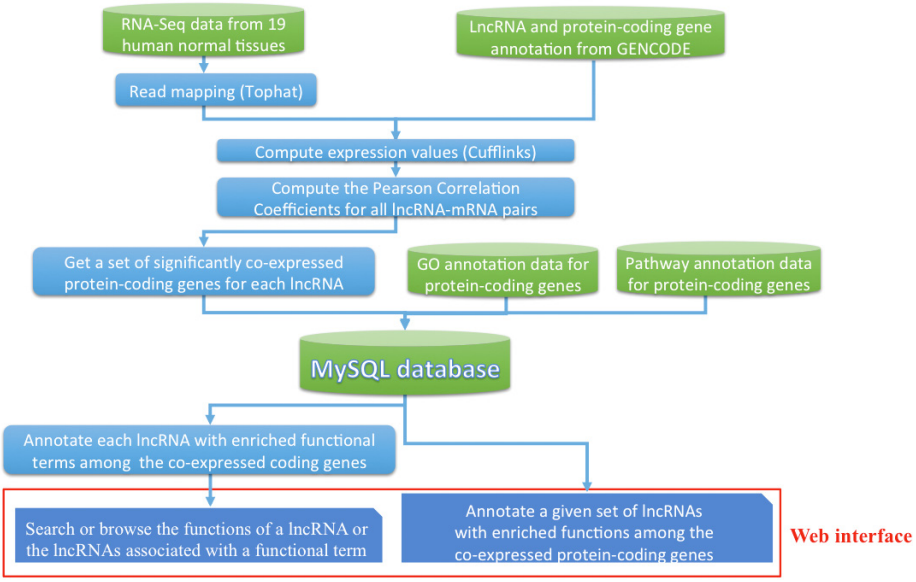

Fig. 1. Overview of the lncRNA2Function

Network Propagation Reveals Novel Genetic Features Predicting Drug Response of Cancer Cell Lines

Jiguang Wang*, Judith Kribelbauer*, and Raul Rabadan**

Department of Biomedical Informatics and Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032 USA
rabadan@dbmi.columbia.edu

Translating data derived from cancer genomes into personalized cancer therapy is a holy grail of computational biology. An important, yet challenging, question in this undertaking is to relate features of tumor cells to clinical outcomes of anticancer drugs. Recent progress in large pharmacogenomic studies has provided a wealth of data about cancer cell lines, indicating that many genetic and gene expression candidates might predict the drug response of cancer cells [1-3]. Unfortunately, most of the predicted features lack underlying mechanisms and are not consistent with our prior knowledge [4].

To address this question, we have developed a new method, named dNetFS, to prioritize gene expression features, as well as genetic features of cancer cell lines, that predict drug response by integrating genomic/pharmaceutical data, protein-protein interaction networks, and prior knowledge of drug-targets interaction with the techniques of network propagation. Compared with previous methods, dNetFS is better than other simple network-based methods and dramatically improves the accuracy of prediction of traditional correlation-based methods by means of cross-validation analysis. Our dNetFS software will be available upon request.

By applying dNetFS in the study of an inhibitor of Insulin-like Growth-Factor-Receptor (IGF1R), BMS-754807 [5], we were able to show that the sensitivity of BMS-754807 could be accurately predicted by the expression levels of some important genes, including proto-oncogene tyrosine-protein kinase Src, and neuroblastoma RAS viral (v-ras) oncogene homolog.

References

1. Barretina, J., et al.: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607 (2012)
2. Basu, A., et al.: An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 154, 1151–1161 (2013)
3. Garnett, M.J., et al.: Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483, 570–575 (2012)
4. Haibe-Kains, B., El-Hachem, N., Birkbak, N.J., Jin, A.C., Beck, A.H., Aerts, H.J., Quackenbush, J.: Inconsistency in large pharmacogenomic studies. Nature 504, 389–393 (2013)

* Contributed equally.
** Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 375, 2014.

Splice Site Prediction Using Support Vector Machine with Markov Model and Codon Information

Dan Wei[1,2], Yin Peng[3,4], Yanjie Wei[2,*], and Qingshan Jiang[2,*]

[1] Institute of Graphics and Image, Hangzhou Dianzi University, Hangzhou 310018, China
[2] Shenzhen Key Lab. for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology,
Chinese Academy of Sciences, Shenzhen 518055, China
[3] Department of Pharmacology, Sun Yat-Sen University, Guangzhou, 510275, China
[4] Department of Center Laboratory, The First Affiliated Hospital to Shenzhen University,
Shenzhen 518035, China

Prediction of donor and acceptor splice sites plays a central role for detecting the gene structure for the eukaryotes. In this paper, we combine the sequence conservativeness and codon usage bias to predict splice sites. Our method is based on SVM with Markov model and codon usage information (MC-SVM). The method first extracts two features of the candidate sequences, including the conserved features described by the probabilistic parameters of the first Markov model (MM1) and the codon usage bias information. Then an F-score based feature selection is used to select the most discriminative features. Finally, MC-SVM applies SVM on the training sequences with sequence-based vectors as input to obtain the SVM prediction model, and uses the model to predict the splice sites of testing sequences.

The proposed method is tested using 10-fold cross-validation on two 1:1 and 1:10 datasets, with all the true splice sites taken from Homo Sapiens Splice Sites Data set (HS3D) and equal/decuple number of false sites randomly selected from the same data set. The evaluation shows that MC-SVM is highly accurate compared to MM1-SVM[1], Reduced MM1-SVM [2] and some other methods [3] in terms of sensitivity, specificity and global accuracy Q^9. Furthermore, ROC curves show that MC-SVM exhibits better overall prediction performance than MM1-SVM, Reduced MM1-SVM and MEM [4] methods for predicting both acceptor and donor sites.

References

1. Baten, A.K.M.A., Chang, B.C.H., Halgamuge, S.K., Li, J.: Splice site identification using probabilistic parameters and SVM classification. BMC Bioinformatics 7(suppl. 5), S15 (2006)
2. Baten, A.K.M.A., Halgamuge, S.K., Chang, B.C.H.: Fast splice site detection using information content and feature reduction. BMC Bioinformatics 9(suppl. 12), S8 (2008)
3. Zhang, Q., Peng, Q., Zhang, Q., Yan, Y., Li, K., Li, J.: Splice sites prediction of human genome using length-variable Markov model and feature selection. Expert Syst. Appl. 37, 2771–2782 (2010)
4. Yeo, G., Burge, C.: Maximum entropy modeling of short sequence motifs with application to RNA splicing signals. J. Comput. Biol. 11(2-3), 377–394 (2004)

* Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 376, 2014.

Similarity Analysis of DNA Sequences Based on Frequent Patterns and Entropy*

Xiaojing Xie[1], Jihong Guan[2], and Shuigeng Zhou[1,**]

[1] School of Computer Science, and Shanghai Key Lab. of Intelligent Information Processing, Fudan University, China
{xiexiaojing,sgzhou}@fudan.edu.cn
[2] Department of Computer Science and Technology, Tongji University, China
jhguan@tongji.edu.cn

Abstract. DNA sequence analysis has been an important research topic in Bioinformatics. Evaluating the similarity between sequences, which is crucial for sequence analysis, has attracted much research effort, and dozens of algorithms and tools have been developed. These methods are based on either alignment or word frequency and geometric representation etc., each of which has its advantage and disadvantage. In this paper, for effectively computing the similarity between DNA sequences, we introduce a novel method based on frequency patterns and entropy to construct representative vectors of DNA sequences. Concretely, each sequence is first divided into blocks of the same length. Then, a modified PrefixSpan [1] algorithm is used to discover the maximal frequent patterns in each block. Finally, with the probabilities of these patterns, the entropy of each block is calculated. The resulting entropies of the blocks constitute the components of the sequence vector. Our method is able to capture fine-granularity information (location and ordering) of DNA sequences, via sequence blocking. As only the maximal frequent patterns are considered, our method is insensitive to noise and sequence rearrangement. Experiments are conducted to evaluate the proposed method, which is compared with two existing methods [2,3]. When testing on the β-globin genes of 11 species and using the results from MEGA as the baseline, our method achieves higher correlation coefficients than the two existing methods.

References

1. Pei, J., et al.: Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Transactions on Knowledge and Data Engineering 16, 1424–1440 (2004)
2. Yu, H.J., Huang, D.S.: Graphical Representation for DNA Sequences via Joint Diagonalization of Matrix Pencil. IEEE Journal of Biomedical and Health Informatics 17, 503–511 (2013)
3. Li, C., et al.: Similarity analysis of DNA sequences based on the weighted pseudo-entropy. Journal of Computational Chemistry 32, 675–680 (2011)

* This work was supported by National Natural Science Foundation of China (NSFC) under grants No. 61173118 and No. 61272380.
** Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 377, 2014.

Exploiting Topic Modeling to Boost Metagenomic Sequences Binning*

Ruichang Zhang[1], Zhanzhan Cheng[1], Jihong Guan[2], and Shuigeng Zhou[1,**]

[1] Shanghai Key Lab. of Intelligent Information Processing, Fudan University, China
{rczhang,chengzhanzhan,sgzhou}@fudan.edu.cn
[2] Department of Computer Science and Technology, Tongji University, China
jhguan@tongji.edu.cn

With the rapid development of high-throughput technologies, researchers can sequence the whole metagenome of a microbial community sampled directly from the environment. The assignment of these sequence reads into different species or taxonomical classes is a vital step for metagenomic analysis, which is referred to as *banning* of metagenomic data.

In this paper, we propose a new method *TM-Cluster* for banning metagenomic reads. First, we represent each metagenomic read as a set of "k-meres" with their frequencies appearing in the read. Then, we employ a probabilistic topic model — the Latent Dirichlet Allocation (LDA) model [1] to the reads, which generates a number of hidden "topics" such that each read can be represented by a distribution vector of the generated topics. Finally, as in the Cluster method TCluster [3], we apply SKWIC [2] — a variant of the classical K-means algorithm with automatic feature weighting mechanism to clustering these reads.

Our method can achieve stable and better overall performance on datasets with from several thousands to millions of reads of a number of species and various relative abundance ratios, compared to existing banning methods including AbundanceBin, MetaCluster 3.0 and MCluster [3]. Analysis on the topic number of LDA model in our method also implies that the topic number hidden in the metagenomic data is related to the species number to some extent. In summary, our experiments indicate that the incorporation of topic modeling can effectively improve the banning performance of metagenomic reads.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Frigui, H., Nasraoui, O.: Simultaneous clustering and dynamic keyword weighting for text documents. In: Survey of Text Mining, pp. 45–72. Springer (2004)
3. Liao, R., Zhang, R., Guan, J., Zhou, S.: A new unsupervised binning approach for metagenomic sequences based on n-grams and automatic feature weighting. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2013), http://doi.ieeecomputersociety.org/10.1109/TCBB.2013.137

* This work was supported by National Natural Science Foundation of China (NSFC) under grants No. 61173118 and No. 61272380.

** Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 378, 2014.

Network-Based Method for Identifying Overlapping Mutated Driver Pathways in Cancer

Hao Wu, Lin Gao*, Feng Li, Fei Song, and Xiaofei Yang

School of Computer Science and Technology, Xidian University,
Xi'an, Shaanxi 710071, China
lgao@mail.xidian.edu.cn, haowu@nwsuaf.edu.cn,
{lifeng_10_28,ronasong,yangxiaofeihe}@163.com

Abstract. Large-scale cancer genomics projects are providing lots of data on genomic, epigenomic and gene expression aberrations in many cancer types [1]. One key challenge is to detect functional driver pathways and to filter out nonfunctional passenger genes in cancer genomics. In this study, we present a network-based method (Net-Dendrix) to detect overlapping driver pathways automatically. This algorithm can directly find driver pathways or gene sets de novo from somatic mutation data utilizing two combinatorial properties, high coverage and high exclusivity [2,3,4], without any prior information. Vandin et al. introduce the Maximum Weight Submatrix Problem to find driver pathways and show that it is an NP-hard problem [2]. To solve it better and reduce the complexity of the problem, we firstly construct gene network based on the approximate exclusivity between each pair of genes using somatic mutation data from lots of cancer patients. Secondly, we present a new greedy strategy to add or remove genes for getting overlapping gene sets with driver mutations according to the properties of high exclusivity and high coverage. To assess the efficiency of Net-Dendrix, we apply it onto simulated data and compare it with Iterative versions of MCMC [2] and RME [4]. Net-Dendrix can obtain the optimal results in less than eight seconds, while Iter-IME can get them in more than 20s and Iter-MCMC can get them in more than 600s. To further verify the performance of Net-Dendrix, we apply it to analyze somatic mutation data from five real biological data sets such as the mutation profiles of 90 glioblastoma tumor samples and 163 lung carcinoma samples. Net-Dendrix detects groups of genes which overlap with known pathways, including P53, RB and PI(3)K signaling pathways. Many gene sets with p-value<1e-3 are found from the somatic mutation data. So Net-Dendrix can detect more biological relevant gene sets. Results show that Net-Dendrix outperforms other algorithms for detecting driver pathways or gene sets.

Keywords: Driver pathway, Network-based method, Somatic mutation, Mutually exclusivity, High coverage.

* Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 379–380, 2014.

References

1. Zhao, J., Zhang, S., Wu, L.-Y., Zhang, X.-S.: Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics 28(22), 2940–2947 (2012)
2. Vandin, F., Upfal, E., Raphael, B.J.: De novo discovery of mutated driver pathways in cancer. Genome Research 22(2), 375–385 (2012)
3. Leiserson, M.D., Blokh, D., Sharan, R., Raphael, B.J.: Simultaneous identification of multiple driver pathways in cancer. PLoS Computational Biology 9(5), e1003054 (2013)
4. Miller, C.A., Settle, S.H., Sulman, E.P., Aldape, K.D., Milosavljevic, A.: Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Medical Genomics 4(1), 34 (2011)

Completing a Bacterial Genome with *in silico* and Wet Lab Approaches

Rutika Puranik[1], Jacob Werner[1], Guangri Quan[2], Rong Zhou[3], and Zhaohui Xu[1,⋆]

[1] Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43403, USA
zxu@bgsu.edu
[2] School of Software, Harbin Institite of Technology, Weihai, Shandong, 264209, China
[3] Department of Mathematics, Yuncheng University, Shanxi, 044000, China

The existence of gaps in draft genome assemblies compromises our ability to take full advantage of genome data. In this study, a pipeline is developed to assemble complete genomes primarily from the next generation sequencing (NGS) data. The input of the pipeline are paired-end Illumina sequence reads, and the output is a high quality complete genome sequence. The pipeline alternates the employment of computational and biological methods in seven steps. It combines the strengths of *de novo* assembly, reference based assembly, customized programming, public databases utilization, and wet lab experimentation.

The application of the pipeline is demonstrated by the completion of a bacterial genome, *Thermotoga* sp. strain RQ7, a potential biohydrogen production strain. Illumina sequencing produced 400 Mb of clean data. Initial assembling with SOAPdenovo [1] and SOAPaligner [2] generated a scaffold of 1,822,593 bp that contained 27 mini gaps, ranging from 1 bp to 3.2 kb, and one big gap of $\sim$ 38 kb. After running through the pipeline, the genome was closed at 1,851,618 bp, with a GC content of 46.13%. The annotation of 63 ORFs were updated, affecting the prediction of many essential cellular processes.

This work distinguishes itself from similar studies [3, 4] due to its multi-phase interactions between computational and biological approaches. The constituting principles and methods are applicable to similar studies of both prokaryotic and eukaryotic genomes.

References

[1] Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al.: De novo assembly of human genomes with massively parallel short read sequencing. Genome Research 20(2), 265–272 (2010)
[2] SOAPaligner, http://soap.genomics.org.cn/soapaligner.html
[3] Nadalin, F., Vezzi, F., Policriti, A.: GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics 13(suppl. 14), S8 (2012)
[4] Xing, Y., Medvin, D., Narasimhan, G., Yoder-Himes, D., Lory, S.: CloG: A pipeline for closing gaps in a draft assembly using short reads. In: 2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), February 3-5, pp. 202–207 (2011)

⋆ Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 381, 2014.

Screening Ingredients from Herbs against Pregnane X Receptor in the Study of Inductive Herb-Drug Interactions: Combining Pharmacophore and Docking-Based Rank Aggregation

Zhijie Cui[1], Hong Kang[1], Kailin Tang[1], Qi Liu[1], Zhiwei Cao[1,2,*], and Ruixin Zhu[1,3,*]

[1] Department of Bioinformatics, Tongji University, Shanghai, P.R. China
[2] Shanghai Center for Bioinformation Technology, Shanghai, P.R. China
[3] School of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian, Liaoning, P.R. China

An issue of integrative medicine about herb-drug interactions has been increasing concerned [1]. Herbal ingredients can activate nuclear receptors to induce drug-metabolizing enzyme and/or transporter expression, which result in altering efficacy and toxicity of co-administrated drugs. This process is called inductive herb-drug interactions [2]. Pregnane X Receptor (PXR) and drug-metabolizing target genes are involved in most of inductive herb-drug interactions. To predict this kind of herb-drug interaction, identifying ligands of nuclear receptors and drug-metabolizing enzyme/transporter could be done respectively. In addition, because drugs and their metabolizing enzymes are well known, the prediction would be simplified to only screen agonists of nuclear receptors. Finally, 421 herbs were collected to build a curated herb-drug interaction database, which records 380 herb-drug interactions including 90 herbs and 230 drugs. This database was used to validate our computational results.

A combinational *in silico* strategy of pharmacophore and docking-based rank aggregation (DRA) was employed to identify PXR's agonists. Firstly, 305 ingredients were screened out from 820 ingredients as candidate agonists of PXR with our pharmacophore model. Secondly, DRA was used to re-rank the result of pharmacophore filtering. Finally, the top 10 ingredients were mapped to 14 herbs, and 5 of these herbs were involved to the reported herb-drug interactions. This study demonstrated that the computational strategy was a promising way to investigate inductive herb-drug interactions.

References

1. Lopez-Picazo, J.J., Ruiz, J.C., Sanchez, J.F., Ariza, A., Aguilera, B., Lazaro, D., Sanz, G.R.: Prevalence and typology of potential drug interactions occurring in primary care patients. The European Journal of General Practice 16(2), 92–99 (2010)
2. Reitman, M.L., Chu, X., Cai, X., Yabut, J., Venkatasubramanian, R., Zajic, S., Stone, J.A., Ding, Y., Witter, R., Gibson, C., et al.: Rifampin's acute inhibitory and chronic inductive drug interactions: experimental and model-based approaches to drug-drug interaction trial design. Clinical Pharmacology and Therapeutics 89(2), 234–242 (2011)

* Corresponding authors.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 382, 2014.

Improving Multiple Sequence Alignment by Using Better Guide Trees

Qing Zhan[1,⋆], Yongtao Ye[2,⋆], Tak-Wah Lam[2], Siu-Ming Yiu[2], Hing-Fung Ting[2,⋆⋆], and Yadong Wang[1,⋆⋆]

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[2] HKU-BGI Bioinformatics Algorithms & Core Technology Research Lab., Department of Computer Science, University of Hong Kong

A commonly used approach for multiple sequence alignment (MSA) is the progressive alignment approach, which first constructs a guide tree that is supposed to capture the phylogenetic relationship of the input sequences, and then aligns the sequences progressively according to the topology of the tree. Previous studies have verified that guide trees are very important to the quality of the resulting alignments. In this work, we investigated how to construct better guide trees for better MSAs. In particular, we study an adaptive guide tree construction method, which was introduced by Ye *et al.*[1] for their MSA tool GLProbs. This method first computes the average percent identity $\overline{\text{PID}}$ of the input sequences, and if $\overline{\text{PID}}$ is small, it explores local information to construct the guide trees, and if $\overline{\text{PID}}$ is large, it focuses on global information. We study whether this adaptive method constructs the best guide trees for GLProbs. We also study whether it can improve the output quality of other MSA tools.

First, we have modified GLProbs to GLProbs-Random and GLProbs-Reference in which the adaptively constructed guide tree used by GLProbs is replaced by a randomly generated tree and an estimated phylogenetic tree (from the reference MSA) respectively. The three columns labeled GLProbs in Table 1 compares their performances with the sum of pairs score (SP): comparing the columns for GLProbs and GLProbs-Random confirmed that the guide trees constructed by the adaptive method do better, and comparing the columns for GLProbs and GLProbs-Reference suggested that the adaptively constructed guide trees are among the best. Next, we have modified five leading tools by replacing their original guide tree construction steps with the adaptive one, and keeping other steps intact. The result in Table 1 shows that all the five modified tools achieved significant improvements, especially when the sequences have low similarity, e.g. ClustalW-Adaptive outperformed its original by 12.3% for sequences of 0-20% similarity.

⋆ Joint first authors.
⋆⋆ Corresponding authors.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 383–384, 2014.

Table 1. Mean SP scores on Benchmark OXBench

	GLProbs			ClustalW		MSAProbs		Probalign		ProbCons		T-Coffee	
similarity	Ref	Ran	Ori	Ada	Ori	Ada	Ori	Ada	Ori	Ada	Ori	Ada	Ori
0%-100%	90.30	90.06	**90.38**	**89.83**	89.44	**90.09**	90.06	**89.99**	89.96	**89.72**	89.68	**89.53**	89.51
0%-20%	**47.63**	46.03	47.33	**48.22**	42.94	**45.14**	44.84	**44.31**	43.57	**45.39**	44.14	**44.88**	43.82

"Ref" denotes using the estimated (maximum likelihood) phylogenetic tree from reference MSA; "Ran" denotes using the randomly generated tree; "Ori" denotes the aligner original version; "Ada" denotes using the adaptive guide tree construction method. Better results are shown in bold.

Reference

1. Ye, Y., Cheung, D.W., Wang, Y., Yiu, S.-M., Zhan, Q., Lam, T.-W., Ting, H.-F.: GLProbs: Aligning Multiple Sequences Adaptively. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, pp. 152–160 (2013)

A Markov Clustering Based Link Clustering Method for Overlapping Module Identification in Yeast Protein-Protein Interaction Networks*

Yan Wang[1,2], Guishen Wang[1], Di Meng[1], Lan Huang[1,**], Enrico Blanzieri[2,**], and Juan Cui[3]

[1] College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education Jilin University, Changchun, China
[2] Department of Information and Communication Technology University of Trento, Povo, Italy
[3] Department of Computer Science and Engineering University of Nebraska at Lincoln, Lincoln, NE, USA
huanglan@jlu.edu.cn, blanzier@disi.unitn.it

Abstract. Previous studies have shown that many overlapping components among the modular structures in protein-protein interaction (PPI) networks reflect common functional components shared by different biological processes. In this paper, we proposed a Markov clustering based Link Clustering (MLC) method to identify the overlapping modular structures in PPI networks. MLC method calculates the extended link similarity that represents the relevance between links, the interactions between proteins, and derives a similarity matrix. It then uses markov clustering to partition the link similarity matrix and obtains overlapping modules in the original network automatically without much parameters and threshold constraints. Our experimental validation on two benchmark networks with known reference classes and the Yeast PPI network, respectively, show that MLC outperforms the original Link Clustering and the classical Clique Percolation Method with higher EQ/ENMI/DR evaluation and better GO enrichment performance. It is particularly interesting that, on Yeast PPI network, MLC also identifies new functional modules in which genes do not show significant correlation among their expressions. Overall, the MLC method has demonstrated promising potentials in identifying the core biological modules or important pathways in different organisms through studying the interplay between functional processes.

Keywords: Overlapping Module, Protein-protein Interaction, Markov Clustering, Link Clustering.

* This work is supported by the Natural Science Foundation of China (61175023), Jilin Innovation Team Project (20121805), the Ph.D. Program Foundation of MOE of China (20120061120106), and the Science-Technology Development Project of Jilin Province of China (20130522111JH, 20130522114JH, 20140101180JC).
** Corresponding authors.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 385, 2014.

Protein Function Prediction: A Global Prediction Method with Multiple Data Sources

Jun Meng[1], Xin Zhang[1], and Yushi Luan[2,*]

[1] School of Computer Science and Technology,
Dalian University of Technology, Dalian, China
[2] School of Life Science and Biotechnology,
Dalian University of Technology, Dalian, China

Multiple types of genomics and proteomics data are being available by heterogeneous high-throughput experiments. As each data source captures only one aspect about proteins' properties, it's necessary and wise to integrate these heterogeneous high-throughput data sources which bring a more complete picture about protein functions. The MS-KNN method [1] shows three data sources for protein function prediction: protein-protein interaction (PPI), gene expression and sequence similarity. Yu [2] proposed a method called TMEC to capture the relationships between pairs of proteins, between pairs of functions, and between proteins and functions. However, many methods predict functions without fully considering the properties of each data source. In order to use these data effectively, we choose appropriate methods for different data source to construct networks and merge those networks.

In this paper, we choose three Yeast data sources for protein function prediction, PPI, gene expression and protein sequence, which roles for function annotation were introduced by MS-KNN. The data sources are downloaded from the Biological General Repository for Interaction Datasets (BioGRID), the *Saccharomyces* Genome Database (SGD) and the MIPS Comprehensive Yeast Genome Database (CYGD), respectively. To calculate the weights between pairs of proteins in a PPI network, edge clustering coefficient [3] is a suitable measure which can evaluate the importance of edges in PPI and describe how close the two proteins are. For gene expression data, Pearson correlation coefficient is a frequently used coefficient to express the degree of linear relationship between two sets of gene expression value. For protein sequence, in consideration of poor similarity between proteins, we extract protein sequence' PseAAC [4] features and calculate the inner product distance between two proteins by the features, instead of sequence homologous similarity-based method. As the efficiency of methods for predicting protein functions from networks depend on the number of non-zero interactions, we sparse dense networks. Therefore, we retain *k*-nearest neighbors for each protein and set the rest to zero for gene expression and protein sequence network. Then, a naïve Bayesian fashion [5] is used to combine the networks. Finally, a global propagation algorithm [6] is designed on the combined

* Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 386–387, 2014.

network, which takes the known function annotations for protein as the sources of 'function flow'. The method considers the global and local network topology.

The experimental results show that the proposed global propagation algorithm by iterating the combined network method has superior over MS-KNN and TMEC with high accuracy of protein function prediction.

References

1. Lan, L., Djuric, N., Guo, Y.H., Vucetic, S.: MS-kNN: protein function prediction by integrating multiple data sources. BMC Bioinformatics 14, S8 (2011)
2. Yu, G., Rangwala, H., Domeniconi, C., Zhang, G.J., et al.: Protein Function Prediction using Multi-label Ensemble Classification. IEEE ACM T. Comput. Bi. 10, 1 (2013)
3. Wang, J., Li, M., Wang, H., Pan, Y.: Identification of essential proteins based on edge clustering coefficient. IEEE ACM T. Comput. Bi. 9, 1070–1080 (2012)
4. Chou, K.C.: Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Struct. Funct. Genet. 43, 246–255 (2001)
5. Von Mering, C., Jensen, L.J., Snel, B., et al.: String: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 33, D433–D437 (2005)
6. Vanunu, O., Magger, O., Ruppin, E., et al.: Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. 6, e1000641 (2010)

A microRNA-Gene Network in Ovarian Cancer from Genome-Wide QTL Analysis

Andrew Quitadamo, Frederick Lin, Lu Tian, and Xinghua Shi

Department of Bioinformatics and Genomics
College of Computing and Informatics
University of North Carolina at Charlotte
Charlotte NC 28223, USA
{aquitada,flin8,ltian,x.shi}@uncc.edu

Ovarian cancer is the most deadly reproductive cancer in women. A better understanding of the biological mechanisms of ovarian cancer is needed for earlier diagnosis and more effective treatment. Differential microRNA(miRNA) expression and miRNA/mRNA dysregulation have been associated with ovarian cancer. Whole-genome miRNA and mRNA sequencing provides a new prospective to study these aberrations for their associations with ovarian cancer.

In this study, we perform a genome-wide QTL analysis between miRNA and gene expression in ovarian cancer, using data from The Cancer Genome Atlas (TCGA) [1]. The results from such QTL analysis provided a network new of the relationship between miRNA and gene expression. We found that all of the identified miRNAs were reported previously to be associated with different diseases, and particularly, the majority of these miRNAs were shown to be associated with ovarian cancer. Our results replicated several cancer genes [2], and provided a list of candidate cancer genes as well. In summary, we showed that our integrative analysis would help understand the molecular mechanism of disease manifestation and progression, and eventually result in better prognosis, diagnosis and treatment of ovarian cancer.

Keywords: Cancer genomics, ovarian cancer, microRNAs, RNA sequencing, Quantitative Trait Loci (QTL) analysis.

References

1. Cancer Genome Atlas Research Network: Integrated genomic analyses of ovarian carcinoma. Nature 474(7353), 609–615 (2011)
2. Atlas of Genetics and Cytogenetics in Oncology and Haematology, http://atlasgeneticsoncology.org/Genes/Geneliste.html

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 388, 2014.

K-Profiles Nonlinear Clustering

Kai Wang[1] and Tianwei Yu[2,*]

[1] Department of Mathematics and Computer Science,
Emory University, Atlanta, Georgia, USA
[2] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA

Modern technologies such as microarray, deep sequencing, liquid chromatography–mass spectrometry (LC-MS) etc make it possible to measure the expression levels of thousands of genes/proteins simultaneously to unravel important biological processes. Detecting nonlinear relationships are most useful in the context of exploratory knowledge discovery from large biological datasets, when data structure itself is not yet well understood. Nonlinear relations, which were mostly unutilized in contrast to linear correlations, are prevalent in high-throughput data. In many cases, it can model biological relationships more precisely and reflect critical patterns in the biological systems.

Clustering is usually taken as the first step towards elucidating hidden patterns and understanding the mass of data. However, no single clustering algorithm tops all performance charts due to its built-in biases on datasets [1]. Well-defined relationship/distance measurement and cluster profiles play crucial roles in the process. Using the general dependency measure, Distance based on Conditional Ordered List (DCOL) that we introduced before [2], we designed the nonlinear K-profiles clustering method, which can be seen as the nonlinear counterpart of the K-means clustering algorithm with statistical testing incorporated to remove prevalent noise in biological data. It not only outperformed our previous General Dependency based Hierarchical Clustering (GDHC) algorithm and the traditional K-means algorithm in our simulation studies, but also showed much improved computational efficiency in contrast with GDHC. Real data analysis showed its capability to detect novel nonlinear patterns in high-throughput data. It will be discussed in detail in this talk.

References

1. D'Haeseleer, P.: How does gene expression clustering work? Nat. Biotechnol. 23(12), 1499–1501 (2005)
2. Yu, T., Peng, H., Sun, W.: Incorporating Nonlinear Relationships in Microarray Missing Value Imputation. IEEE/ACM Trans. Comput. Biol. Bioinform. 8(3), 723–731 (2011)

* Corresponding author.
1518 Clifton Rd NE, Rm 334, Atlanta, GA 30322.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 389, 2014.

Estrogen Induced RNA Polymerase II Stalling in Breast Cancer Cell Line MCF7

Zhi Han[1,2], Lu Tian[3], Jie Zhang[4], Tim Huang[5], Raghu Machiraju[6], and Kun Huang[2,*]

[1] College of Software, Nankai University, Tianjin, China
[2] Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA
[3] Department of Health Policy and Research – Biostatistics, Stanford University, USA
[4] OSU Biomedical Informatics Shared Resource, The Ohio State University, USA
[5] Department of Genetic Medicine, University of Texas Health Science Center, USA
[6] Department of Computer Science and Engineering, The Ohio State University, USA

RNA polymerase II (PolII) stalling is an important phenomenon in gene regulation. This is an important cellular process in response to stress [1]. For the genes with PolII stalling, while the PolII molecules accumulate at their promoter regions, their transcription processes are paused. Here we investigate PolII stalling induced by estrogen in the breast cancer cell line MCF7 by integrating data from ChIP-seq and microarray technologies. We take a rule based approach to identify genes with PolII stalling after 17β-estrodial (E2) treatment in MCF7 cells. E2 treatment activates estrogen receptor which plays important roles in the majority of breast cancers [2]. We use ChIP-seq data for PolII and gene expression microarray data for MCF7 before and after treatment of E2.

Our method includes several main steps: First, we identify genes with enriched PolII binding segment using a signal processing based algorithm we have previously developed. This algorithm identifies both long and short enriched regions from ChIP-seq data [3]. Next, we select genes whose PolII enrichment levels increase at the promoter regions but decrease over the gene bodies after E2 treatment. Finally, we zoom into genes whose expression levels decreased significantly ($p < 0.05$ and mean fold change > 1.5) after E2 treatment. We also apply similar rules to identify genes released from PolII stalling after E2 treatment.

Our method identified 92 genes which satisfy our criteria and demonstrate PolII stalling induced by E2 in MCF7 cell line while only 3 genes show PolII stalling released by E2. Functional analysis of identified genes shows that E2 induced PolII stalling is highly relevant to cancer development pathways. This suggests that E2 treatment potentially can cause a stress response of the breast cancer cells which leads promotion or disruption of cancer related biological functions.

References

[1] Baugh, L.R., et al.: RNA Pol II accumulates at promoters of growth genes during developmental arrest. Science 324, 92–94 (2009)
[2] Fox, E.M., et al.: ERbeta in breast cancer–onlooker, passive player, or active protector? Steroids 73, 1039–1051 (2008)
[3] Han, Z., et al.: A signal processing approach for enriched region detection in RNA polymerase II ChIP-seq data. BMC Bioinformatics 13(suppl. 2), S2 (2012)

* Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 390, 2014.

A Knowledge-Driven Approach in Constructing a Large-Scale Drug-Side Effect Relationship Knowledge Base for Computational Drug Discovery

Rong Xu[1] and QuanQiu Wang[2]

[1] Case Western Reserve University, Cleveland OH 44106, USA
[2] ThinTek, LLC, Palo Alto CA 94306

Introduction. It has been increasingly recognized that similar side effects of seemingly unrelated drugs can be caused by their common off-targets and that drugs with similar side effects are likely to share molecular targets [1]. Therefore, systems approaches to studying side effect relationships among drugs and integration of this drug phenotypic data with drug-related genetic, genomic, proteomic, and chemical data will facilitate drug target discovery and drug repositioning [2]. The availability of a comprehensive drug-side effect (SE) relationship knowledge base is critical for these tasks. Current drug phenotype-driven systems approaches rely exclusively on drug-SE associations extracted from FDA drug labels. However, there exists a large amount of additional drug-SE relationship knowledge in the large body of published biomedical literature. In this study, we present a novel knowledge-driven (KD) approach to automatically extract a large number of drug-SE pairs from 21 million published biomedical abstracts. We systematically analyzed extracted drug-SE pairs in combination with drug-related gene targets, metabolism, pathways, gene expression and chemical structure data. We show that these extracted drug-SE pairs have great potential in drug discovery.

Methods. Our study is based on the two key observations: (1) multiple side effects for a drug are often reported in the same sentences or abstracts; and (2) if a sentence contains a known drug-SE pair, then this sentence is likely to be SE-relevant. Other pairs in this SE-related sentence are likely to be drug-SE pairs. In this study, we used all known drug-SE pairs derived from FDA drug labels as prior knowledge to find SE-related MEDLINE sentences and abstracts, from which many additional drug-SE pairs that have not included in FDA drug labels are then extracted. We compared the KD approach to a support vector machine (SVM)-based approach. The entire experimental process consists of the following steps: (1) Build a local MEDLINE search engine; (2) Develop, evaluate and compare the KD approach to a SVM-based approach; (3) Extract drug-SE pairs from MEDLINE; and (4) Systematically analyze the correlation between drug-associated side effects and drug gene targets, metabolism genes, chemical similarity, and disease indications. For the text corpus, we used 21,354,075 syntactically parsed MEDLINE records (119,085,682 sentences).

Results. First, we used known drug-SE associations derived from FDA drug labels as prior knowledge to automatically find SE-related sentences and abstracts. We then extracted a total of 49,575 drug-SE pairs from MEDLINE sentences and 180,454

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 391–392, 2014.

pairs from abstracts. On average, the KD approach has achieved a precision of 0.335, a recall of 0.509, and an F1 of 0.392, which is significantly better than a SVM-based machine learning approach (precision: 0.135, recall: 0.900, F1: 0.233) with a 73.0% increase in F1 score. Through integrative analysis, we demonstrate that the higher-level phenotypic drug-SE relationships reflect lower-level genetic, genomic, and chemical drug mechanisms. In addition, we show that the extracted drug-SE pairs can be directly used in drug repositioning.

Conclusions. In summary, we automatically constructed a large-scale drug phenotype relationship knowledge, which in combination with other genetic, genomic and chemical data resources, can have great potential in computational drug discovery.

References

1. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., Bork, P.: Drug target identification using side-effect similarity. Science 321(5886), 263–266 (2008)
2. Hurle, M.R., Yang, L., Xie, Q., Rajpal, D.K., Sanseau, P., Agarwal, P.: Computational drug repositioning: From data to therapeutics. Clinical Pharmacology and Therapeutics (2013)

Systems Biology Approach to Understand Seed Composition

Ling Li[1,*], Wenxu Zhou[2], Manhoi Hur[1], Joon-Yong Lee[1], Nick Ransom[1], Cumhur Yusuf Demirkale[3], Zhihong Song[2], Dan Nettleton[3], Mark Westgate[4], Vidya Iyer[5], Jackie Shanks[5], Eve Syrkin Wurtele[1], and Basil J. Nikolau[2,*]

[1] Department of Genetics, Development and Cell Biology
[2] Department of Biochemistry, Biophysics and Molecular Biology
[3] Department of Statistics
[4] Department of Agronomy
[5] Department of Chemical and Biological Engineering,
Iowa State University, Ames, Iowa 50011, USA
{liling,dimmas}@iastate.edu

Abstract. As the propagule that ensures the dissemination of plants, seeds also support human activity as one of the major products of agriculture. The biochemical storage reserves that are deposited within the seed during its development chemically fall into three general categories, proteins, oils and carbohydrates. The seed reserves are biosynthesized by the programmed expression of a metabolic network during seed development. In most commercial lines of soybean grown in the Midwestern states of the US, seeds are composed of 40% protein, 20% oil, 15% soluble carbohydrates, and 15% fiber (http://www.asa-europe.org/SoyInfo/composition_e.htm). There is considerable knowledge concerning the basic biochemical processes by which imported carbon and nitrogen is converted to the final products, protein, oil and carbohydrate. However, there is a great deal to be learned concerning the molecular, biochemical and genetic mechanisms that regulate this complex metabolic network. Recent developments in genomics have provided the catalogue of genes that would be required for this process. We have taken advantage of combined metabolomics and transcriptomics technologies to identify the global developmental and biochemical transcriptomics network, and ultimately determine structure and composition of the mature seed. Also, we have coupled this with bioinformatics and metabolic flux analyses to gain insights as to the biochemical programs that determine soybean seed development. For this purpose, we have developed Plant & Microbial Metabolomics Resource (PMR, http://www.metnetdb.org/pmr), a platform to empower the use of metabolomics data in the development of hypotheses concerning the organization and regulation of metabolic networks, and MetNet systems biology platform (http://www.metnetdb.org) for plant 'omics, a web-based framework which enables interactive visualization of metabolic and regulatory networks. This combination of genetic resources, high-throughput experimental data and bioinformatic analyses has revealed sets of specific genes, genetic perturbations and mechanisms, and metabolic changes that are associated seed composition during soybean seed development.

Keywords: *Glycine max*, Evans, seed development, gene expression, metabolomic change, seed composition, PMR, MetNet.

* Corresponding authors.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 393, 2014.

Prediction of the Cooperative *cis*-regulatory Elements for Broadly Expressed Neuronal Genes in *Caenorhabditis Elegans*

Chen Xu and Zhengchang Su

Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, 351 Bioinformatics Building, 9201 University City Blvd, Charlotte, NC 28223, USA

How cell types are derived by gene regulatory programs is a fundamental problem in developmental biology. The nervous system of the *Caenorhabditis elegans* (*C. elegans*) provides an excellent model to gain a good understanding of this problem. It has been shown that common structural features shared by diverse types of neurons in *C. elegans*, such as axons, dendrites and synapses, etc. are determined by a set of genes broadly expressed in neuronal cells, known as pan-neuronal genes [1]. However, so far very little is known about the transcriptional regulatory mechanisms of these genes. In this research, we found that two of our identified putative motifs tend to concur in the vast majority of upstream intergenic regions of the pan-neuronal genes [2], and hence are likely to be a *cis*-regulatory module (CRM). Interestingly, this module is also widely distributed in the whole *C. elegans* genome and is highly conserved between *C. elegans* and *Caenorhabditis briggase* (*C. briggase*). In addition to the general control by a common CRM, the pan-neuronal genes may rely on other *cis*-regulatory motifs in order to further specify their functions. We found that some identified motifs were harbored by different subsets of pan-neuronal genes which could be significantly related to different functions respectively. These results suggest that the pan-neuronal gene features are defined by the collective usage of the two regulatory mechanisms. Our computational results should provide some hints of *cis*-regulatory mechanisms of the pan-neuronal genes and a useful guide to experimental validations.

References

[1] Hobert, O., Carrera, I., Stefanakis, N.: The molecular and gene regulatory signature of a neuron. Trends in Neurosciences 33, 435–445 (2010)

[2] Ruvinsky, I., Ohler, U., Burge, C.B., Ruvkun, G.: Detection of broadly expressed neuronal genes in C. Elegans. Developmental Biology 302, 617–626 (2010)

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 394, 2014.

Improving the Mapping of the Smith-Waterman Sequence Database Search Algorithm onto CUDA GPUs*

Chao-Chin Wu[1], Liang-Tsung Huang[2], Lien-Fu Lai[1], and Yun-Ju Li[1]

[1] Department of Computer Science and Information Engineering
National Changhua University of Education, Changhua 500, Taiwan
{ccwu,lflai}@cc.ncue.edu.tw, icecloud6666@gmail.com
[2] Department of Biotechnology, Mingdao University, Changhua 523, Taiwan
larry@mdu.edu.tw

Sequence alignment is one of the most important methodologies in the field of computational biology. The most widely used sequence alignment algorithm may be the Smith-Waterman algorithm because of its high sensitivity of sequence alignment even though it has higher time complexity of algorithm [1]. To enable the Smith-Waterman algorithm produce exact results in a reasonably shorter time, much research has been focusing on using various high-performance architectures to accelerate the processing speed of the algorithm. In particular, it becomes a recent trend to use the emerging accelerators and many-core architectures to run the Smith-Waterman algorithm.

Modern general-purpose (Graphics Processing Units) GPUs are not only powerful graphics engines, but also highly parallel programmable processors [2]. Today's GPUs use hundreds of parallel processor cores executing tens of thousands of parallel threads to rapidly solve large problems, now available in many PCs, laptops, workstations, and supercomputers. Because of the availability and the popularity, GPUs have been used to implement the Smith-Waterman algorithm, where CUDASW++ is the leading reaseach that provides the fast, publicly available, solution to the exact Smith-Waterman algorithm on commodity hardware. CUDASW++ 3.0 is the latest version, which couples CPU and GPU SIMD instructions and carries out concurrent CPU and GPU computations [3].

This paper focuses on how to improve CUDASW++, especially for short query sequences. We observe that the shared memory in each streaming multiprocessor is not fully utilized in CUDASW++. Therefore, the execution flow of the Smith-Waterman algorithm is rearranged to fully utilize the shared memory for reducing the amount of slow global memory access. We have added our approach to CUDASW++ 2.0 and run experiments on nVIDIA Tesla C1060 and C2050. Experimental results demonstrate that our approach outperforms CUDASW++.

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. Journal of Molecular Biology 147, 195–197 (1981)
2. CUDA GPUs, https://developer.nvidia.com/cuda-gpus
3. Liu, Y., Wirawan, A., Schmidt, B.: CUDASW++ 3.0: Accelerating Smith-Waterman Protein Database Search by Coupling CPU and GPU SIMD Instructions. BMC Bioinformatics 14, 117 (2013)

* This work is supported by the contract, NSC102-2221-E-018-024.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 395, 2014.

Isomorphism and Similarity for 2-Generation Pedigrees

Haitao Jiang[1], Guohui Lin[2], Weitian Tong[2], Daming Zhu[1], and Binhai Zhu[3]

[1] School of Computer Science and Technology,
Shandong University, Jinan, Shandong, China
{htjiang,dmzhu}@sdu.edu.cn
[2] Department of Computing Science, University of Alberta,
Edmonton, Alberta T2G 2E6, Canada
{weitian,guohui}@ualberta.ca
[3] Department of Computer Science, Montana State University,
Bozeman, MT 59717, USA
bhz@cs.montana.edu

In this paper, we follow the work by Kirkpatrick *et al.* [4] to consider the isomorphism and similarity problems for the simplest (unlabeled) pedigrees — 2-generation pedigrees, where the isomorphism and similarity problems are both studied. We show that the isomorphism problem is GI-hard (GI — Graph Isomorphism) even for 2-generation pedigrees. If the 2-generation pedigrees are monogamous (i.e., each individual at level-1 can mate with exactly one partner) then the isomorphism testing problem can be solved in polynomial time.

Subsequently, we relax the similarity measure for two general 2-generation pedigrees by using the minimum number of isomorphic $\langle i, j\rangle$-families which they can be decomposed into. Here, an $\langle i, j\rangle$-family is a sub-family of a couple with i female children and j male children. It turns out that this can be formulated as a Minimum Common Integer Pair Partition (MCIPP) problem, generalizing the NP-complete Minimum Common Integer Partition (MCIP) problem [1]. We then exploit a new property of the optimal solution for MCIPP, and show that MCIPP is Fixed-Parameter Tractable [2,3].

Acknowledgments. This research is partially supported by NSF of China under grant 60928006, 61070019 and 61202014, by NSF of Shandong Province under grant ZR2012FQ008, and by NSERC of Canada.

References

1. Chen, X., Liu, L., Liu, Z., Jiang, T.: On the minimum common integer partition problem. ACM Trans. on Algorithms 5(1) (2008)
2. Downey, R., Fellows, M.: Parameterized Complexity. Springer (1999)
3. Flum, J., Grohe, M.: Parameterized Complexity Theory. Springer (2006)
4. Kirkpatrick, B., Reshef, Y., Finucane, H., Jiang, H., Zhu, B., Karp, R.: Comparing pedigree graphs. J. of Computational Biology 19(9), 998–1014 (2012)

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 396, 2014.

VFP: A Visual Tool for Predicting Gene-Fusion Base on Analyzing Single-end RNA-Sequence

Ye Yang[1,2] and Juan Liu[1,*]

[1] School of Computer, Wuhan University, Wuhan, Hubei, China
[2] Military Economy Academy, Wuhan, Hubei, China
liujuanjp@163.com

Gene fusion is a key factor in sarcomas, lymphomas, leukemias and so on. In recent years, some fusion detection algorithms were published to search the fusion by the data produced on next generation sequencing platform, Such as TopHat-Fusion[1], FusionHunter[2], FusionSeq[3]. But these algorithms have some common defects, such as the limitation of the operating system, the confusion of the parameter setting and so on. In order to help biologist to quickly discover the target of the treatment, we have developed VFP to predict gene-fusion from single-end RNA-sequencing reads. VFP employs seed index strategy and octal encoding operations for sequence alignments and uses several rules to score and filter the potential fusion genes. We tested VFP by a simulated dataset and two real datasets and found that VFP can detect known and novel fusions in lymphoma and melanoma datasets.

There are many extensions and modifications of VFP, some of them will be mentioned in the talk.

References

1. Kim D, Salzberg SL: TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biology. 12(2011),R72.
2. Li Y, et al: FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics. 27(2011), 1708−1710 .
3. Sboner A. et al: FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biology. 11(2010),R104.

* Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 397, 2014.

A Novel Method for Identifying Essential Proteins from Active PPI Networks

Qianghua Xiao[1,2], Xiaoqing Peng[1], Fangxiang Wu[1,3], and Min Li[1]

[1] School of Information Science and Engineering,
Central South University, Changsha 410083, China
[2] School of Mathematics and Physics,
University of South, HengYang 421001, China
[3] Division of Biomedical Engineering,
University of Saskatchewan, Saskatoon, SK, S7N 5A9, Canada

Essential proteins are vital for cellular survival and development. Identifying essential proteins is very important for helping us understand the way in which a cell works. Rapid increase of available protein-protein interaction (PPI) data has made it possible to detect protein essentiality at the network level. A series of centrality measures have been proposed to discover essential proteins based on the PPI networks. However, the PPI data obtained from large scale, high-throughput experiments generally contain false positives. It is insufficient to use original PPI data to identify essential proteins.

In this paper, we firstly adopt a dynamic model-based method to filter noisy data from time-course gene expression profiles. Second, a threshold of each protein is calculated from a threshold function of σ, the protein is active at a time point if its expression level is higher than the threshold. Two proteins are regarded as co-expression if they are all active at the same time point. Finally, an active PPI network is constructed by combining gene expression data with PPI data.

The classical centrality measures, like Degree centrality(DC), Local Average Connectivity Centrality (LAC), Edge Clustering Coefficient (NC), Betweenness Centrality (BC), Closeness Centrality (CC), and Subgraph Centrality (SC), are methods which can be applied to identify essential proteins based on network topology. These centrality measures are redefined and performed to identify essential proteins on the active PIN. The experimental results on yeast network show that the performance of centrality measures to identify essential proteins are considerably improved based on the active PPI network, compared with original PPI network, in terms of the number of identified essential proteins in top %k percentage and a jackknife methodology. At the same time, the results also indicate that most essential proteins are active.

Acknowledgement. This work is supported in part by the National Natural Science Foundation of China under Grant No.61370024, No.61232001, and No.61379108, the Program for New Century Excellent Talents in University (NCET-12-0547), Science and Technology Plan Projects of Science and Technology Bureau of Hengyang City (grant 2013KJ29).

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 398, 2014.

RAUR: Re-alignment of Unmapped Reads with Base Quality Score

Xiaoqing Peng[1], Zhen Zhang[1], Qianghua Xiao[1,2], and Min Li[1]

[1] School of Information Science and Engineering,
Central South University, Changsha 410083, China
[2] School of Mathematics and Physics, University of South,
HengYang 421001, China

In recent years, with the emergencies of next-generation genome sequencing technologies, many software tools have been developed to efficiently and accurately align short reads to the reference genome. However, for most alignment tools, the edit distances or the allowed mismatches are limited, thus some reads cannot be mapped if their mismatches in any hits exceed the allowable differences.

Some trimmed-like strategies appear in some alignment programs and try to handle the problem. For example, Bowtie2 and BWA-MEM in BWA can perform local read alignment for long reads by maximizing the alignment score. However, the false positive sites are also introduced, since the maximum alignment score can't make sure that high quality bases are involved.

In this article, we propose a method (RAUR) to re-align the unmapped reads. A trimming strategy used in RAUR is to figure out the longest and most confident and informative segment of a read based on base quality score. RAUR can be applied on any alignment tool if there are reads which can't be aligned by it. It adopts an iterative progress to trim the unmapped reads until the reads can be confidently mapped or can't be mapped in any progress. To evaluate the performance of RAUR, we apply RAUR on the simulated reads and real reads of human genome with different lengths by comparing with BWA, Bowtie2, and SOAP2 with different settings. From the *precision* and *alignment rate*, we can find out that RAUR can improve the *alignment rate*s greatly, especially for long reads, while the *precision*s are still comparative with the original alignments. RAUR proposes a new insight for re-aligning unmapped reads, which can contribute to the downstream analysis.

Acknowledgement. This work is supported in part by the National Natural Science Foundation of China under grant nos. 61232001, 61379108, and 61370172, Hunan Provincial Innovation Foundation For Postgraduate (CX2013B070), and Science and Technology Plan Projects of Science and Technology Bureau of Hengyang City (grant 2013KJ29).

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 399, 2014.

PIGS: Improved Estimates of Identity-by-Descent Probabilities by Probabilistic IBD Graph Sampling

Danny S. Park[1], Yael Baran[2], Farhad Hormozdiari[3], and Noah Zaitlen[1]

[1] University of California San Francisco, San Francisco CA 94143, USA
[2] Tel Aviv University, Tel Aviv, Israel
[3] University of California Los Angeles, Los Angeles CA 90095, USA

Identifying segments of the genome that are identical-by-descent (IBD) between individuals is a fundamental concept in genetics. IBD data are used in numerous applications including demographic inference, heritability estimation, and disease loci mapping. Therefore, the identification of IBD segments from genome-wide genotyping studies, and more recently sequencing studies, has important implications for studies of complex human traits.

Current methods for detecting IBD fall into two categories: multiway or pairwise. Multiway methods detect IBD over multiple haplotypes simultaneously and leverage the clique structure of true IBD. Although powerful, these approaches are generally computationally expensive since the number of potential IBD relationships at a locus is $O(2^{h(h-1)/2})$, where h is the number of haplotypes. As a result, many state-of-the-art methods estimate the probability of IBD between pairs of haplotypes independently [1]. The result is a much more efficient method but at the expense of loss of power when detecting smaller IBD segments (<1 centimorgans) [2].

We develop a hybrid approach (PIGS), which combines the computational efficiency of pairwise methods and the power of multiway methods. It leverages the IBD clique structure to simultaneously compute the probability of IBD conditional on all pairwise estimates. We show over extensive simulations, that PIGS yields a substantial increase in the number of identified small IBD segments. We observed a 95% increase in the total number of identified IBD segments of 0.5 centimorgans and a 40% increase in identified IBD segments across all sizes.

Given the substantial improvement in the number of identified IBD segments from our method, we expect that the approach will greatly facilitate the discovery of new loci from IBD-based disease association studies.

References

1. Browning, B.L., Browning, S.R.: Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194(2), 459–471 (2013)
2. He, D.: IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. Bioinformatics 29(13), i162–i170 (2013)

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 400, 2014.

Clustering PPI Data through Improved Synchronization-Based Hierarchical Clustering Method

Xiujuan Lei[1,2,*], Chao Ying[3], Fang-Xiang Wu[4], and Jin Xu[5]

[1] School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710062, China
[2] School of Electronics Engineering and Computer Science,
Peking University, Beijing, 100871, China
xjlei@snnu.edu.cn
[3] School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710062, China
nbsschao@163.com
[4] Division of Biomedical Engineering, University of Saskatchewan,
Saskatoon, SK S7N 5A9, Canada
faw341@mail.usask.ca
[5] School of Electronics Engineering and Computer Science,
Peking University, Beijing, 100871, China
jxu@pku.edu.cn

Abstract. Clustering algorithm is the main method to identify function module of protein-protein interaction (PPI) network, but traditional methods have advantages and corresponding drawbacks. The synchronization-based hierarchical clustering (SHC) algorithm was improved in this paper. Firstly, the PPI data was preprocessed via spectral clustering (SC) method which transforms the high-dimensional similarity matrix into a low dimension matrix. Then the SHC algorithm is used to perform clustering. In SHC algorithm, hierarchical clustering are achieved by enlarging the local neighborhood distance of objects synchronizing continuously, while the hierarchical search is very difficult to find the optimal local neighborhood distance of synchronizing and the efficiency is not high. So the glowworm swarm optimization (GSO) algorithm was adopted to determine the optimal threshold of the local neighborhood distance of synchronization automatically. The algorithm is tested on the PPI dataset. The results show that the improve algorithm is better than the traditional algorithms in *precision*, *recall* and *f-measure*.

Keywords: Protein-Protein Interaction network, glowworm swarm optimization algorithms, synchronization-based hierarchical clustering, spectral clustering algorithm.

1 Introduction

Protein-protein interaction (PPI) network is an important research field in the bioinformatics. Identifying the function of protein complex is critical for understanding

* Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 401–403, 2014.

disease mechanisms, diagnosis and therapy. Recently there are a large number of clustering methods applied to discover modules in PPI network, but they suffered some degree of shortcomings. The cluster number of spectral clustering(SC)[1] must be predefined and the clustering algorithm is sensitive to noise data. Synchronization-based hierarchical clustering(SHC)[2] is very difficult to find the optimal local neighborhood distance of synchronizing and the efficiency is not high. In order to overcome these defects, we preprocess the PPI data via transforming the high-dimensional similarity matrix into a low dimension matrix inspired by spectral clustering, then the SHC algorithm is used to perform clustering and glowworm swarm optimization algorithm(GSO)[3] is used to find the optimal threshold of local neighborhood distance of synchronizing.

2 Methods and Results

The pretreatment process of spectral clustering need to construct a similarity matrix A, follow as:

$$A_{ij} = \begin{cases} w\dfrac{\left|N_i \text{ I } N_j\right|+1}{\min(\left|N_i\right|,\left|N_j\right|)} + (1-w)\dfrac{\sum_{k\in I_{i,j}} w(i,k)\cdot\sum_{k\in I_{i,j}} w(j,k)}{\sum_{s\in N_i} w(i,s)\cdot\sum_{t\in N_j} w(j,t)}, & i \neq j \\ 0, & i = j \end{cases} \tag{1}$$

Then constructing Laplacian matrix L on the basis of matrix A. Matrix X consist of matrix L's eigenvector the first three eigenvalue corresponding, normalize matrix X. Clustering the processed data make use of synchronization-based hierarchical clustering and replace hierarchical search by GSO algorithm, which to find the optimal threshold of local neighborhood distance in SHC.

Dynamic synchronous model applied in this paper follows as:

$$x_i(t+1) = x_i(t) + \frac{1}{\left|N_\varepsilon(x(t))\right|}\sum_{y\in N_\varepsilon(x(t))}\sin(y_i(t)-x_i(t)) \tag{2}$$

The objective function(*fval*) of GSO algorithm follows as:

$$fval = \sum_{i=1}^{x}\left\{(2\cdot m_{H_i}/(n_{H_i}\cdot(n_{H_i}-1)))^{\rho}\cdot\left(\sum_{u,v\in H_i, w_{u,v}\in W} w_{u,v}\Big/\sum_{v\in H_i, w_{v,k}\in W} w_{v,k}\right)^{1-\rho}\right\} \tag{3}$$

The maximum value of *fval* corresponds to the optimal local neighborhood distance of synchronization.

Hierarchical clustering method based on dynamic synchronous model proposed the concept of neighborhood closures, reducing the running time of synchronization clustering algorithm. Meanwhile the efficiency and accuracy of the algorithm is improved by using GSO algorithm to determine the optimal choice thresholds of local neighborhood distance of synchronizing. The *recall* and *precision* values of the new algorithm are improved and the anti-noise ability is better compared with SC and SHC algorithms, but the time complexity is relatively higher and still need to be decreased.

Acknowledgment. This paper is supported by the National Natural Science Foundation of China (61100164, 61173190), Scientific Research Start-up Foundation for Returned Scholars, Ministry of Education of China ([2012]1707) and the Fundamental Research Funds for the Central Universities, Shaanxi Normal University (GK201402035, GK201302025).

References

[1] Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems, vol. 14, pp. 849–856. MIT Press, Cambridge (2001)

[2] Huang, J., Kang, J., Qi, J., Sun, H.: A hierarchical clustering method based on a dynamic synchronization model. Science China: Information Science 43(5), 599–610 (2013)

[3] Krishnanand, K.N., Ghose, D.: Glowworm swarm optimization: a new method for optimizing multi-modal functions. International Journal of Computational Intelligence Studies 1(1), 93–119 (2009)

Order Decay in Transcription Regulation in Type 1 Diabetes*

Shouguo Gao[1], Shuang Jia[2,3], Martin J. Hessner[2,3], and Xujing Wang[1,**]

[1] Bioinformatics and Systems Biology Core, Systems Biology Center, National Heart, Lung and Blood Institute, NIH, Bethesda, MD 20892, USA
[2] The Max McGee National Research Center for Juvenile Diabetes, Department of Pediatrics at the Medical College of Wisconsin and the Children's Research Institute of the Children's Hospital of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin, 53226, USA
[3] The Human and Molecular Genetics Center, The Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin, 53226, USA
xujing.wang@nih.gov

In this study, using type 1 diabetes (T1D) as a model system, we investigate two outstanding challenges in the development of molecular signatures of complex traits: the incorporation of gene interaction structure in signature definition, and the integration of multiple Omics data types to refine signature. The T1D data consists of our previously published transcription profiles in control *peripheral blood mononuclear cells* (*PBMC*) induced by sera of 142 human subjects from unrelated healthy controls (uHC), and 3 T1D family cohorts: recent onset (RO-T1D), and healthy siblings of probands that are at high (HRS) or low (LRS) genetic risk for T1D. First both weighted and non-weighted co-expression networks were separately constructed in each cohort and were compared. Several network measures, including edge weight and degree distribution, Shannon's entropy, the λ-coefficient, and h-index, were determined. We found that overall the co-expression networks induced by the RO-T1D cohort are significantly weaker, exhibiting a broad spectrum loss of order and control. More specifically, all T1D family cohorts induced more active and orderly transcription coordination among the innate immunity genes, consistent with our previous report of them sharing a heightened innate inflammatory state. On the other hand, higher coordination of the adaptive immunity genes was only induced by the LRS cohort, potentially explaining their low risk for disease. All the network measures also pointed to the same story, and additionally the importance of the innate immunity genes in determining the transcriptome state of the T1D family cohorts. Next, we integrated the protein-protein interaction (PPI) and the transcriptomic co-expression networks, and focused specifically on the smallest functional units of PPI, the protein complexes (PC). A PC is considered active in a cohort, if its co-expression network is percolated. We found that the RO-T1D cohort activated a significant less number of PC than the others. Overall, whether it is co-expression or protein interaction networks, the four cohorts show striking differences and can be clearly discriminated based on network structural measures. In contrast, gene expression levels alone, without the consideration of underlying interaction networks, could barely differentiate the cohorts. In summary these findings demonstrate the advantage network based metrics in defining molecular signatures.

** Corresponding author.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 404, 2014.

Simulated Regression Algorithm for Transcriptome Quantification

Adrian Caciula[1], Olga Glebova[1], Alexander Artyomenko[1], Serghei Mangul[2], James Lindsay[3], Ion I. Măndoiu[3], and Alex Zelikovsky[1]

[1] Georgia State University, Atlanta GA, 30303, USA
[2] University of California, Los Angeles CA, 90095, USA
[3] University of Connecticut, Storrs CT, 06269, USA

RNA-Seq is a cost-efficient high-coverage powerful technology for transcriptome analysis. We propose a novel algorithm for transcriptome quantification from RNA-seq data (*SimReg*) which uses regression to find transcript frequencies for which the simulated read counts match the observed read counts.

SimReg first aligns the reads to existing transcript library and then counts the equivalent reads, i.e., reads aligned to the same set of transcripts. The bipartite graph with vertices corresponding to transcripts and read classes is split into connected components which can be treated independently. For each component we simulate high coverage reads and estimate $D_{\mathcal{R},\mathcal{T}} = \{d_{r,t}\}$, where $d_{r,t}$ is the portion of reads from transcript $t \in \mathcal{T}$ belonging to read class $r \in \mathcal{R}$.

Initial transcript frequencies are estimated by minimizing the squared deviation between observed read class frequency $O_{\mathcal{R}} = \{o_r\}$ and expected read class frequency $E_{\mathcal{R}} = D_{\mathcal{R},\mathcal{T}} \times F_{\mathcal{T}}$, where $F_{\mathcal{T}} = \{f_t\}$ are the portions of reads emitted by transcripts. The squared deviation is minimized by the following quadratic program: $\sum_{r\in\mathcal{R}} \left(\sum_{t\in\mathcal{T}} d_{r,t} f_t - o_r\right)^2 \to \min \,|\, \sum_{t\in\mathcal{T}} f_t = 1$ and $f_t \geq 0$.

Next *SimReg* repeatedly updates the frequency estimates by (1) simulating reads according to current estimates $F_{\mathcal{T}}$, (2) finding deviation between simulated and observed reads, $\Delta_{\mathcal{R}} = S_{\mathcal{R}} - O_{\mathcal{R}}$, (3) obtaining corrected read frequencies $C_{\mathcal{R}} = O_{\mathcal{R}} - \Delta_{\mathcal{R}}/2$, and (4) updating estimated transcript frequencies $F_{\mathcal{T}}$ based on corrected read class frequencies $C_{\mathcal{R}}$.

We tested *SimReg* on several test cases using simulated human RNA-Seq data. Experiments on synthetic RNA-seq datasets show that the proposed method improves transcriptome quantification accuracy compared to previous methods. The results show better correlation compared with currently best method *RSEM* [1].

Reference

1. Li, B., Dewey, C.: Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC Bioinformatics 12(1), 323 (2011)

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, p. 405, 2014.

Author Index