MLAB
Sapporo
2012

2012 Sapporo Workshop on Machine Learning and Applications to Biology

August 6 - 7, 2012, Sapporo, Hokkaido

supported by

Creative Research Institution (CRIS), Hokkaido University

Steering Committee

Hiroshi Mamitsuka (Kyoto University)

Ichigaku Takigawa (Hokkaido University)

Preface

MLAB (Workshop on Machine Learning and Applications to Biology) Sapporo 2012 is held in Sapporo, Japan on August 6-7, 2012. This meeting is the first attempt in Japan to hold a scientific meeting on machine learning development and particularly that for research subjects in biology. The meeting consists of invited talks devoted by twenty prominent researchers who are actively working on machine learning, bioinformatics and/or systems biology. This volume contains the program of the meeting and the abstracts of the twenty talks presented at MLAB Sapporo 2012. The talks cover a wide variety of topics in machine learning and bioinformatics. Typical examples in machine learning are feature selection, divergence estimation, kernels, label propagation, link prediction and frequent subsequence/subgraph mining, while sequence mutation detection, inferring drug-target interactions and predicting RNA folds are instances in the biology side. We believe that this meeting leads to lively discussion among the participants, inspiring new perspectives, which contributes to promoting research in machine learning as well as that for biology.

All information about MLAB Sapporo 2012 is available at

http://www.cris.hokudai.ac.jp/takigawa/mlab2012/

To conclude, we would like to explicitly acknowledge and thank the tremendous support by the Creative Research Institution (CRIS) of Hokkaido University and the kind cooperation by the Institute for Chemical Research of Kyoto University.

Conference organizers
Hiroshi Mamitsuka
Ichigaku Takigawa

Program Schedule: DAY 1

August 6 (Monday)

Each talk will be 35min long

8:30am - 9:00am	**Registration and Welcome**
9:00am - 11:20am	**Session 1**
	01: **Masashi Sugiyama (Tokyo Institute of Technology)** Recent Advances in Divergence Estimation: Theory, Algorithm, and Application
	02: **Shanfeng Zhu (Fudan University)** Efficient Semi-Supervised MEDLINE Document Clustering
	03: **Masayuki Karasuyama (Kyoto University)** Label Propagation through Graph-based Feature Reconstruction
	04: **Timothy Hancock (Kyoto University)** Imposing Network Structures on Feature Selection on Experimental Data
11:20am - 2:00pm	**Lunch Break**
2:00pm - 3:45pm	**Session 2**
	05: **Hisashi Kashima (The University of Tokyo)** Link Prediction Methods for Bioinformatics
	06: **Canh Hao Nguyen (Kyoto University)** Latent Feature Kernels for Link Prediction
	07: **Motoki Shiga (Toyohashi University of Technology)** Efficient Semi-Supervised Learning on Multiple Graphs
3:45pm - 4:00pm	**Break**
4:00pm - 5:45pm	**Session 3**
	08: **Jean-Philippe Vert (Mines ParisTech)** Structured Feature Selection for Genomic Data
	09: **Ichigaku Takigawa (Hokkaido University)** Learning Sparse Linear Models over Subgraph Indicators
	10: **Yasuo Tabei (JST ERATO)** Space-Efficient Multibit Tree for Large-Scale Chemical Fingerprint Searches
5:45pm -	**Group Photo**
6:30pm -	**Banquet**

Program Schedule: DAY 2

August 7 (Tuesday)

Each talk will be 35min long

9:00am - 11:20am	**Session 1**
	11: **Jun Sese (Tokyo Institute of Technology)** Gene Expression Analysis in Polyploid Species using Next-Generation Sequencer
	12: **Yuichi Shiraishi (The University of Tokyo)** An Empirical Bayesian Framework for Mutation Detection from Cancer Genome Sequencing Data
	13: **Hiroto Saigo (Kyushu Institute of Technology)** Learning from Treatment History to Predict Response to Anti-HIV Therapy
	14: **Yoshihiro Yamanishi (Kyushu University)** Machine Learning Methods to Analyze and Infer Drug-Target Interaction Networks
11:20am - 2:00pm	**Lunch Break**
2:00pm - 3:45pm	**Session 2**
	15: **Satoshi Morinaga (NEC)** Factorized Asymptotic Bayesian Inference for Learning Latent Variable Models
	16: **Marco Cuturi (Kyoto University)** Distances and Kernels on Discrete Structures
	17: **Hiroki Arimura (Hokkaido University)** Efficient Enumeration of Bounded-Size Subtrees in a Tree and Its Application to Tree Mining with Proximity Constraint
3:45pm - 4:00pm	**Break**
4:00pm - 5:45pm	**Session 3**
	18: **Koji Tsuda (AIST)** Fast Similarity Search with Succinct Trees
	19: **Kengo Sato (Keio University)** Simultaneous Aligning and Folding of RNA Sequences via Dual Decomposition
	20: **Atsuyoshi Nakamura (Hokkaido University)** Frequent Pattern Mining for Families of Dispersed Repeats in DNA Sequences
5:45pm -	**Closing**

Abstracts

01	**Masashi Sugiyama**	*Recent Advances in Divergence Estimation: Theory, Algorithm, and Application*
02	**Shanfeng Zhu**	*Efficient Semi-Supervised MEDLINE Document Clustering*
03	**Masayuki Karasuyama**	*Label Propagation through Graph-based Feature Reconstruction*
04	**Timothy Hancock**	*Imposing Network Structures on Feature Selection on Experimental Data*
05	**Hisashi Kashima**	*Link Prediction Methods for Bioinformatics*
06	**Canh Hao Nguyen**	*Latent Feature Kernels for Link Prediction*
07	**Motoki Shiga**	*Efficient Semi-Supervised Learning on Multiple Graphs*
08	**Jean-Philippe Vert**	*Structured Feature Selection for Genomic Data*
09	**Ichigaku Takigawa**	*Learning Sparse Linear Models over Subgraph Indicators*
10	**Yasuo Tabei**	*Space-Efficient Multibit Tree for Large-Scale Chemical Fingerprint Searches*
11	**Jun Sese**	*Gene Expression Analysis in Polyploid Species using Next-Generation Sequencer*
12	**Yuichi Shiraishi**	*An Empirical Bayesian Framework for Mutation Detection from Cancer Genome Sequencing Data*
13	**Hiroto Saigo**	*Learning from Treatment History to Predict Response to Anti-HIV Therapy*
14	**Yoshihiro Yamanishi**	*Machine Learning Methods to Analyze and Infer Drug-Target Interaction Networks*
15	**Satoshi Morinaga**	*Factorized Asymptotic Bayesian Inference for Learning Latent Variable Models*
16	**Marco Cuturi**	*Distances and Kernels on Discrete Structures*
17	**Hiroki Arimura**	*Efficient Enumeration of Bounded-Size Subtrees in a Tree and Its Application to Tree Mining with Proximity Constraint*
18	**Koji Tsuda**	*Fast Similarity Search with Succinct Trees*
19	**Kengo Sato**	*Simultaneous Aligning and Folding of RNA Sequences via Dual Decomposition*
20	**Atsuyoshi Nakamura**	*Frequent Pattern Mining for Families of Dispersed Repeats in DNA Sequences*

Recent Advances in Divergence Estimation: Theory, Algorithm, and Application

Masashi Sugiyama*, sugi@cs.titech.ac.jp

* Department of Computer Science, Tokyo Institute of Technology

Abstract

Estimation of a divergence between two probability distributions from two sets of data samples is a fundamental challenge in statistical machine learning. An accurate divergence estimator can be used for various purposes such as two-sample homogeneity testing [1], [2], unsupervised change-point detection in time series [3], [4], and semi-supervised class balance estimation under class-prior change [5]. A divergence estimator can also be used also for measuring independence among random variables via mutual information estimation. Such an independence measure can be used for independence testing [6], feature selection [7], [8], dimensionality reduction [9], [10], independent component analysis [11], canonical dependence analysis [12], clustering [13], [14], object matching [15], and causal inference [16]. In this talk, I give an overview of recent theoretical and algorithmic advances in divergence estimation and show its applications [17].

References

[1] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.

[2] T. Kanamori, T. Suzuki, and M. Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720, 2012. to appear.

[3] Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2012.

[4] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. Technical Report 1203.0453, arXiv, 2012.

[5] M. C. Du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In *Proceedings of 29th International Conference on Machine Learning (ICML2012)*, Edinburgh, Scotland, Jun. 26–Jul. 1 2012.

[6] M. Sugiyama and T. Suzuki. Least-squares independence test. *IEICE Transactions on Information and Systems*, E94-D(6):1333–1336, 2011.

[7] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52, 2009.

[8] W. Jitkrittum, H. Hachiya, and M. Sugiyama. Feature selection via ℓ_1-penalized squared-loss mutual information. submitted.

[9] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*. to appear.

[10] M. Yamada, G. Niu, J. Takagi, and M. Sugiyama. Computationally efficient sufficient dimension reduction via squared-loss mutual information. In C.-N. Hsu and W. S. Lee, editors, *Proceedings of the Third Asian Conference on Machine Learning (ACML2011)*, volume 20 of *JMLR Workshop and Conference Proceedings*, pages 247–262, Taoyuan, Taiwan, Nov. 13-15 2011.

[11] T. Suzuki and M. Sugiyama. Least-squares independent component analysis. *Neural Computation*, 23(1):284–301, 2011.

[12] M. Karasuyama and Sugiyama. Canonical dependency analysis based on squared-loss mutual information. submitted.

[13] M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya. On information-maximization clustering: Tuning parameter selection and analytic solution. In L. Getoor and T. Scheffer, editors, *Proceedings of 28th International Conference on Machine Learning (ICML2011)*, pages 65–72, Bellevue, Washington, USA, Jun. 28–Jul. 2 2011.

[14] M. Kimura and M. Sugiyama. Dependence-maximization clustering with least-squares mutual information. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 15(7):800–805, 2011.

[15] M. Yamada and M. Sugiyama. Cross-domain object matching with model selection. In G. Gordon, D. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS2011)*, volume 15 of *JMLR Workshop and Conference Proceedings*, pages 807–815, Fort Lauderdale, Florida, USA, Apr. 11-13 2011.

[16] M. Yamada and M. Sugiyama. Dependence minimizing regression with model selection for non-linear causal inference under non-Gaussian noise. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010)*, pages 643–648, Atlanta, Georgia, USA, Jul. 11–15 2010. The AAAI Press.

[17] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge, UK, 2012.

Efficient Semi-Supervised MEDLINE Document Clustering

Jun Gu*, Wei Feng[†], Jia Zeng[‡], Hiroshi Mamitsuka[§], and Shanfeng Zhu*
zhusf@fudan.edu.cn

* School of Computer Science, Fudan University, China
[†] School of Computer Science and Technology, Tianjin University, China
[‡] School of Computer Science and Technology, Soochow University, China
[§] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

Abstract

Combining multiple information for biomedical document clustering is a subject of intense research. For example, recently the performance of document clustering was enhanced by using both content and MeSH (Medical Subject Heading) semantic information, which were however linearly combined [1]. The simple linear combination could be ineffective, because its representation space is too limited to combine multiple information sources considering the difference in their reliability. To relax this problem, we propose a new semi-supervised clustering method, SSNCut, which incorporates positive (must-link) and negative (cannot-link) constraints in terms of the cost function of spectral learning with normalized cut. We apply SSNCut to MEDLINE document clustering, reasonably assuming that document pairs with high semantic similarities have positive constraints to be in the same cluster and those with low similarities have negative constraints to be in different clusters. Experimental results with various 100 datasets of MEDLINE records show that SSNCut outperformed a linear combination method, being statistically significant.

Acknowledgment

This project was supported by National Nature Science Foundation of China (No. 60903076).

References

[1] SF. Zhu, J. Zeng, and H. Mamitsuka. Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. Bioinformatics 25(15):1944-1951, 2009.

Label Propagation through Graph based Feature Reconstruction

Masayuki Karasuyama*, and Hiroshi Mamitsuka*
{karasuyama, mami}@kuicr.kyoto-u.ac.jp

* Bioinformatics Center, Institute for Chemical Research, Kyoto University

Abstract

Graph-based semi-supervised learning (e.g., [1]–[4]) has received significant attention in machine learning community. We consider a possible and typical framework in this learning paradigm that can be a two-step procedure (which hereafter we call TSP): 1) generating the adjacency matrix from given data, where instances correspond to nodes of an undirected graph, and 2) estimating node labels by using the adjacency matrix. We particularly focus on the first step, i.e. estimating edge weights from given data, because a well-accepted idea in graph-based semi-supervised learning is so-called *manifold assumption*, in which instances with the same labels should be connected with larger weights, meaning that methods in this paradigm are highly affected in performance by edge weights.

Edge weights are very important but given labeled instances are usually limited, making hard to obtain optimal weights. However many existing methods have ignored this essential problem, and in reality rather simple heuristic strategies have been employed to obtain edge weights. A simple and standard one is to use the Gaussian kernel with the width parameter to which some location estimator (e.g., mean or median) for the distance between connected instances is assigned. Another more sophisticated approach is *local reconstruction* [5]–[9] that exploit similar idea to *locally linear embedding* (LLE) [10], which can be free from tuning parameters but noise-sensitive, by which resultant weights cannot be optimized well.

Labeled nodes can be sparse, while all instances have feature vectors. Thus we propose a new method based on *feature vector propagation* (FVP), in which information of feature vectors is propagated on a graph where edge weights are defined as a parameterized similarity function over node pairs. Edge weights are then estimated by optimizing the prediction accuracy of FVP which can be easily evaluated by standard cross-validation, since all nodes including unlabeled nodes have feature vectors. A clear advantage is that FVP is irrelevant to the number of labeled nodes which is in reality limited, causing a serious problem in graph-based semi-supervised learning.

We can show that this approach leads to an objective function similar to that of LLE, when we consider *leave-one-out* (LOO) cross-validation as an evaluation measure. However a significant difference is that edge weights are optimized in a non-parametric manner in LLE but by parameters of the similarity function in FVP. This difference leads to the following two advantages of FVP:

- FVP can alleviate the problem of over-fitting for weight estimation, due to the optimization of the similarity function parameters. In fact, we observed in our experiments that FVP is more robust against noisy data than LLE, in which it is hard to capture locally linear structures on the manifold [11].
- In FVP, edge weights are defined as a parameterized similarity function, by which each resultant weight also represents the similarity of each corresponding node pair. This is very reasonable for many label propagation methods, applicable to the second step of TSP.

In addition, it is possible to make FVP have the original local reconstruction property of LLE. Due to all these properties, it is highly reasonable to use FVP to any dataset which can be assumed to follow the manifold assumption.

Several methods for the second step of TSP attempt to estimate edge weights at the same time, by regarding the entire problem as the graph hyper-parameter optimization. For example, one method minimizes the entropy of estimated node label scores [2] and another similar one maximizes the smoothness of those scores [12]. However, they have a problem of degenerate solutions, for which heuristics are proposed but the validity of the heuristics is unclear. Other methods are based on model selection in supervised learning, such as *kernel target alignment* (KTA) [13], *marginal likelihood maximization* [14] and Leave-one-out (LOO) cross-validation [15]. However a serious problem is that these model selection criteria rely on labeled examples, which are usually limited. Thus the advantages of FVP over such hyper-parameter optimization methods are that: 1) the objective function of FVP does not degenerate, 2) FVP is irrelevant to label information and 3) FVP can be applied to multi-class problems.

We emphasize that any hyper-parameter optimization method cannot satisfy all the three points. In addition, such methods are basically for the second step, meaning that they can be combined with that for the first step, for which FVP was proposed.

We will present our basic framework and show experimental results which demonstrate the effectiveness of FVP through both of synthetic and real datasets.

Acknowledgment

This work has been supported in part by KAKENHI (24300054) of MEXT, Japan.

References

[1] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In C. E. Brodley and A. P. Danyluk, editors, *Proceedings of the 18th International Conference on Machine Learning*, pages 19–26. Morgan Kaufmann, 2001.

[2] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning: From gaussian fields to gaussian processes. In T. Fawcett and N. Mishra, editors, *Proceedings of the 20th Annual International Conference on Machine Learning*. AAAI Press, 2003.

[3] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[4] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[5] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering*, 20:55–67, 2008.

[6] S. I. Daitch, J. A. Kelner, and D. A. Spielman. Fitting a graph to vector data. In *Proceedings of the 26th International Conference on Machine Learning*, pages 201–208, New York, NY, USA, 2009. ACM.

[7] H. Cheng, Z. Liu, and J. Yang. Sparsity induced similarity measure for label propagation. In *IEEE 12th International Conference on Computer Vision*, pages 317–324. IEEE, 2009.

[8] W. Liu, J. He, and S.-F. Chang. Large graph construction for scalable semi-supervised learning. In *Proceedings of the 27th International Conference on Machine Learning*, pages 679–686. Omnipress, 2010.

[9] E. Elhamifar and R. Vidal. Sparse manifold clustering and embedding. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 55–63, 2011.

[10] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[11] J. Chen and Y. Liu. Locally linear embedding: a survey. *Artificial Intelligence Review*, 36:29–48, 2011.

[12] H. H. Shin, N. J. Hill, and G. Rätsch. Graph based semi-supervised learning with sharper edges. In *Proceedings of the 17th European Conference on Machine Learning*, pages 401–412, Berlin, Heidelberg, 2006. Springer-Verlag.

[13] X. Zhu, J. Kandola, Z. Ghahramani, and J. Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1641–1648. MIT Press, Cambridge, MA, 2005.

[14] A. Kapoor, Y. A. Qi, H. Ahn, and R. Picard. Hyperparameter and kernel learning for graph based semi-supervised classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 627–634. MIT Press, Cambridge, MA, 2006.

[15] X. Zhang and W. S. Lee. Hyperparameter learning for graph based semi-supervised learning algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1585–1592. MIT Press, Cambridge, MA, 2007.

Imposing Network Structures On Feature Selection On Experimental Data

Timothy Hancock*, Hiroshi Mamitsuka*
{timhancock, mami}@kuicr.kyoto-u.ac.jp
*Institute for Chemical Research, Kyoto University

Abstract

Networks have become a common place to represent the relationship structure across many variables. For small numbers of variables, networks provide an intuitive model of the structure present within a dataset. However, as the size of the network model increases its representative power diminishes. In an effort to maintain the effectiveness of large network models, feature selection algorithms are employed to extract the relevant structure which is related to a specific phenomena. Currently, the requirement for accurate network feature selection algorithms is essential as the sizes and complexities of the known networks continue to grow.

Supervised classification algorithms are also commonly used as network feature selection methods. The size of the network structures has lead to the development of network regularization algorithms such as the overlap group lasso [1]. Network regularized models place the optimization emphasis on identifying important network nodes and the edge structure is treated as correlation or group structure which is accounted for by the addition of a penalization function. The optimization task then exploits the sparsity assumption made by regularized methods to identify the minimum set of network nodes required to optimize classification performance. When the network structure is large and noisy the sparsity assumption made by penalized methods is appropriate it will enforce the selection of the minimal set of features required for accurate classification. However some networks in biology, such as metabolic networks, are known to possess highly coordinated responses to external phenomena. These responses potentially activate large sections of the network.

In highly correlated environments a related class of models, ensemble methods such as bagging and boosting, are known to perform well. Ensemble methods seek to represent the structure of a large complex dataset through a combination of small models which are built on a subset of important dataset features. In this research we observed an analogous idea to ensemble methods within factorized network probability distributions. Based upon this similarity we propose two novel optimization algorithms, Boosted Expectation Propagation (BEP) and Boosted Message Passing (BMP) [2], [3]. Neither BEP nor BMP assume a sparse solution, but instead seek a weighted average of all network features where the weights are used to emphasize all features which are useful for classification. In this research we focus on applying BEP and BMP to real world networks. Furthermore, we investigate the similarity in selected features and performance between BEP and BMP and network regularized models. We first conduct simulation experiments to highlight the effect of correlated network features on all models. Then we compare model performances on two different types of biological networks, metabolic networks and protein-protein interaction (PPI) networks using microarray data. Our results on these real world networks confirm our simulation results and show that to extract features from correlated networks the assumption of sparsity will adversely effect classification accuracy and feature selection ability.

Acknowledgments

Timothy Hancock was partly supported by BIRD. Hiroshi Mamitsuka was supported in part by BIRD of Japan Science and Technology Agency (JST) and in part by KAKENHI (24300054) of MEXT, Japan.

References

[1] L. Jacob, G. Obozinski, and J. P. Vert, "Group lasso with overlap and graph lasso," *ICML*, 2009.

[2] T. Hancock and H. Mamitsuka, "Boosted optimization for network classification," *Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), Sardinia, Italy*, 2010.

[3] T. Hancock and H. Mamitsuka, "Boosted Network Classifiers for Local Feature Selection," *IEEE Transactions on Neural Networks (Conditional Accept)*,2012.

Link Prediction Methods for Bioinformatics

Hisashi Kashima
kashima@mist.i.u-tokyo.ac.jp

Graduate School of Information Science and Technology, The University of Tokyo

Abstract

In this talk, I introduce two topics related to biological network prediction using machine learning techniques.

The first part is about simultaneous inference of biological networks of multiple species [1]. Most of the existing supervised methods for biological network inference work on each of the networks individually based only on intra-species information such as gene expression data. However, it will be more effective to use genomic data and cross-species evolutionary information from different species simultaneously, rather than to use the genomic data alone. Our semi-supervised learning method called Link Propagation simultaneously infers biological networks of multiple species based on genome-wide data and evolutionary information. This method was successfully applied to simultaneous reconstruction of three metabolic networks of C. elegans, H. pylori, and S. cerevisiae, based on gene expression similarities and amino acid sequence similarities. The Link Propagation method is further extended to simultaneous prediction of different networks in one species [2].

The second part is not directly related to bioinformatics, but I introduce our recent approach for predicting relationships among multiple heterogeneous objects [3], which we believe useful also in biological network prediction. I introduce a multinomial relation prediction method which is robust to data sparsity, which transforms each multinomial relation instance into a set of binomial relations between the objects and the multinomial relation involving the objects, and then an extension of a low-dimensional embedding technique is applied to the binomial relations. The formulation results in a generalized eigenvalue problem guaranteeing global optimal solutions. Experiments with various real-world social web service datasets demonstrate that the proposed method is more robust against data sparseness compared to several existing methods which can only find sub-optimal solutions.

Acknowledgment

I would like to thank the collaborators including Danushka Bollegara, Tsuyoshi Kato, Nozomi Nori, Masashi Sugiyama, Koji Tsuda, and Yoshihiro Yamanishi.

References

[1] H. Kashima, Y. Yamanishi, T. Kato, M. Sugiyama, and K. Tsuda. Simultaneous Inference of Biological Networks of Multiple Species from Genome-wide Data and Evolutionary Information: A Semi-supervised Approach. 25(22):2962–2968, 2009.

[2] H. Kashima, T. Kato, Y. Yamanishi, M. Sugiyama, and K. Tsuda. Link Propagation: A Fast Semi-supervised Learning Algorithm for Link Prediction. In *Proc. 2009 SIAM SIAM Conference on Data Mining (SDM'09)*, Sparks, Nevada, 2009.

[3] N. Nori, D. Bollegara, and H. Kashima. Multinomial Relation Prediction in Social Data: A Dimension Reduction Approach. In *Proc. 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, Toronto, Ontario, Canada, 2012.

Latent Feature Kernels for Link Prediction

Canh Hao Nguyen, and Hiroshi Mamitsuka
{canhhao,mami}@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University

Abstract

Problem Predicting new links in networks is a problem of interest in biological networks. One wishes to computationally suggest potential links to speed up the experimental processes and to explain the biological mechanisms. Most of the methods use only the information on the networks' entities such as sequences of the proteins or gene expression profiles. These methods usually suffer from the information redundancy problem. However, in many applications, network structures (topologies) are known to have patterns, such as in social networks and biological networks. We wish to utilize network structures to build models for links in biological networks for link prediction.

Structures of *similarity networks* are usually used in social networks. However, the problem is that in the networks of protein-protein interactions (PPI), a link does not have similarity semantics, making them *nonsimilarity networks*. A PPI is caused by an interaction between two domains from the two proteins. However, the knowledge of domains and domain-domain interactions is incomplete. Therefore, we wish to incorporate the domain-domain interaction knowledge in an implicit way and latent feature models are for this situation.

Method We derive latent feature models for links from network structures. The idea is that a link is generated from a pair of latent features. Since the features are not observed, many methods generate features explicitly such as Indian Buffet Processes (IBP). However, these methods are too computationally demanding for the sizes of data, which are usually of thousands of nodes.

We propose to use the supervised learning approach for its scalability to real datasets. We formulate the link prediction problem as a problem of classifying pairs of nodes to be in the link class (interaction or regulation) or not. We explicitly model a network as a graph with latent features. The adjacency matrix of the graph is modeled by products of a latent feature matrix and a feature interaction matrix. From the model, we build a similarity matrix of nodes in the graph using the ratios of common neighbors. We propose to use Support Vector Machines by turning the similarity matrix into a kernel matrix. We build a kernel matrix of pairs of nodes based on the nodes' kernel using pairwise kernels, named *latent feature kernels*. After that, a SVM is learnt to classify the classes. More details are in [1].

Result We apply our method to predict link on PPI networks of yeast and fruit-fly, and another nonsimilarity network of gene regulations of E.coli. The following conclusions are drawn from experiments.

- In comparison with the methods for similarity networks, our methods gave higher performances. It means that the networks are nonsimilarity networks and our method was able to take that fact into account.
- In comparison with the methods that generate features explicitly such as IBP, our method was much faster. Our method scaled to real data sizes while IBP only scaled to networks of less than one thousand nodes.
- In comparison with the methods using information of sequences of proteins, our methods gave higher performances. It means that there is much redundant information in sequences for the task, and our method was able to extract the relevant information from only network structures.

Acknowledgment

C.H.N and H.M. have been supported in part by KAKENHI (24300054) of MEXT, Japan.

References

[1] Nguyen, C.H. and Mamitsuka, H. Kernels for Link Prediction with Latent Feature Models. *Proceedings of ECML PKDD 2011, LNAI 6912*, 2011.

Efficient Semi-Supervised Learning on Multiple Graphs

Motoki Shiga* and Hiroshi Mamitsuka†
shiga@cs.tut.ac.jp, mami@kuicr.kyoto-u.ac.jp

* Department of Computer Science and Engineering, Toyohashi University of Technology
† Bioinformatics Center, Institute for Chemical Research, Kyoto University

Abstract

We address an issue of semi-supervised learning on multiple graphs over which informative subgraphs are distributed. In this problem setting, we have two inputs: (1) a set of labeled and unlabeled examples, and (2) multiple (and different) graphs (or networks), in which each node corresponds to an example and each edge indicates similarity between two linked corresponding nodes. We then attempt to estimate labels of unlabeled examples by using connectivity of given multiple graphs. We can find this problem setting in a lot of applications. An example is gene annotation, in which some genes are already annotated but most of all genes are still functionally unknown, despite all experimental efforts in molecular biology . However, the recent development of experimental techniques in biology provides us with a rich number of gene networks, which can be derived from different types of experiments, such as cDNA microarray, protein-protein interactions and ChIP-on-Chip, which are all gene affinity networks but capture different features of gene similarity. Thus the purpose in this application is to assign functions of unlabeled genes by using known genes and multiple graphs.

We present a powerful, time-efficient approach for this problem by combining soft spectral clustering with label propagation for multiple graphs [1]. We propose an efficient approach for our problem setting, allowing to consider the situation in which locally informative subgraphs can be overlapped with each other in each given graph. Our approach is based on a combination of soft spectral clustering [2] and label propagation [3]. More concretely, we first decompose each given graph into eigenvector spectra by which multiple graphs, representing latent clusters, can be generated. We perform this spectral decomposition over all given graphs and then estimate graph weights over all generated graphs by using efficient label propagation algorithm for multiple graphs. Finally we can integrate all subgraphs with their weights by using label propagation for multiple graphs . The final integrated graph can easily assign labels to unknown nodes by their connected edges, while it would be not so easy by using the simply integrated graph.

We empirically demonstrate the performance and efficiency of our proposed approach by using a variety of datasets, including synthetic and real datasets. Experimental results clearly showed the significant performance achievements, the time-efficiency over a large network and the output comprehensibility of our method.

Acknowledgment

This work has been supported in part by BIRD, JST of Japan, KAKENHI (20700269, 24300054), MEXT of Japan, and the Collaborative Research Program of Institute for Chemical Research, Kyoto University (grant #2012-24).

References

[1] M. Shiga and H. Mamitsuka, Efficient Semi-Supervised Learning on Locally Informative Multiple Graphs, *Pattern Recognition*, vol.45, issue 3, pages 1035–1049, 2012.

[2] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, A survey of kernel and spectral methods for clustering, *Pattern Recognition*, vol. 41 issue 1, pages 176–190, 2008.

[3] T. Kato, H. Kashima, and M. Sugiyama, Robust label propagation on multiplenetworks, *IEEE Transactions on Neural Networks*, vol. 20 issue 1, pages 35–44, 2009.

Structured Feature Selection for Genomic Data

Jean-Philippe Vert*
Jean-Philippe.Vert@mines.org

* Mines ParisTech, Institut Curie, INSERM

Abstract

Feature selection for classification or regression is important with high-dimensional genomic data to identify predictive and prognostic biomarkers. It is however a difficult task due to the very large number of features compared to the number of samples usually available, and the large correlation between features.

In this talk I will discuss different approaches to extend classical feature selection methods, such as the Lasso regression, to perform structured feature selection, where we impose some constraints on the features that should be selected such as promoting the selection of genes which are connected to each other on a given gene network. I will particularly focus on the group lasso, an extension of the Lasso which allows to jointly select predefined groups of variables, and will discuss two extensions of the group lasso. First, I will present a new fast method for multiple change-point detection in multidimensional signals, which boils down to a group Lasso regression problem and allows to detect frequent breakpoint location in DNA copy number profiles with millions of probes [1], [2]. Second, I will discuss the latent group lasso, an extension of the group lasso when groups can overlap, which enjoys interesting consistency properties and can be helpful for structured feature selection in high-dimensional gene expression data analysis for cancer prognosis [3], [4].

Acknowledgment

Joint work with Kevin Bleakley, Guillaume Obozinski and Laurent Jacob.

References

[1] J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. NIPS 2010.

[2] K. Bleakley and J.-P. Vert. The group fused Lasso for multiple change-point detection. Technical report HAL-00602121, June, 2011.

[3] L. Jacob, G. Obozinski and J.-P. Vert. Group Lasso with Overlaps and Graph Lasso. ICML 2009.

[4] G. Obozinski, L. Jacob and J.-P. Vert. Group Lasso with Overlaps: the Latent Group Lasso approach. Technical report HAL:inria-00628498, October, 2011.

Learning Sparse Linear Models over Subgraph Indicators

Ichigaku Takigawa* and Hiroshi Mamitsuka†
takigawa@cris.hokudai.ac.jp, mami@kuicr.kyoto-u.ac.jp

* Creative Research Institution, Hokkaido University
† Institute for Chemical Research, Kyoto University

Abstract

In this talk, we discuss statistical inference with linear models over graphs. For given n pairs of graph g_i and response value y_i, $(g_1, y_1), (g_2, y_2), \ldots, (g_n, y_n)$, the target problem is to model the response y to input graph g as $y = \mu(g)$ with a model function μ based on these training samples $\{(g_i, y_i)\}_{i=1}^n$. We assume given graphs g_i are connected, undirected, and labeled as seen in quantitative structure-activity relationship (QSAR) models over molecular graphs. This problem would appear when we try to relate some biological entities encoded as "graphs" to their measurement values obtained from experiments.

Let function μ be a linear model based on countably infinite number of subgraph features $x_1, x_2, \cdots$

$$\mu(g \,|\, \beta, \beta_0) := \beta_0 + \sum_{i=1}^{\infty} \beta_i I(x_i \subseteq g), \quad \beta = (\beta_1, \beta_2, \ldots)$$

where x_i are all possible connected graphs, and $I(x_i \subseteq g)$ denotes a zero-one indicator of subgraph x_i in g that takes 1 for x_i included in g, and 0 otherwise. This also means we characterize the response $y = \mu(g)$ by zero-one explanatory variables indicating the existence of subgraph features $x_1, x_2, \ldots$ in the input graph g.

Then, along the line with previous inspiring works [1]–[3], we present a generalization that can solve the following class of statistical problems with the above linear model μ:

$$\min_{\beta, \beta_0} \sum_{i=1}^{n} L\big(y_i, \mu(g_i \,|\, \beta, \beta_0)\big) + \lambda_1 \|\beta\|_1 + \frac{\lambda_2}{2} \|\beta\|_2^2$$

where $\lambda_1, \lambda_2 \geqslant 0$ and L is a twice differentiable loss function. The penalty term is of elastic-net-type balancing 1-norm (lasso-type) and 2-norm (ridge-type) regularizers. The 1-norm regularizer is a sparsity inducer, and as a consequence, we can naturally select a small number of subgraph features contributing most to the minimization. During the estimation process, these subgraphs are automatically searched from all possible subgraphs, and the coefficients for non-contributing ones remain all zero. As a non-trivial example of this framework, we also show some results on penalized logistic regression over graphs.

Acknowledgment

This work is supported by Creative Research Institution, Hokkaido University, and MEXT/JSPS KAKENHI (No. 23710233, No. 24300054).

References

[1] T. Kudo, E. Maeda, and Y. Matsumoto. An application of boosting to graph classification. In *Advances in Neural Information Processing Systems 17, Neural Information Processing Systems (NIPS'04)*, pages 729–736, Vancouver, Canada, 2004.

[2] K. Tsuda. Entire regularization paths for graph data. In *Proceedings of the 24th International Conference on Machine Learning (ICML'07)*, pages 919–926, Corvalis, USA, 2007.

[3] H. Saigo, S. Nowozin, T. Kadowaki, T. Kudo, and K. Tsuda. gBoost: a mathematical programming approach to graph classification and regression. *Journal of Machine Learning*, 75(1):69–89, 2009.

Space-Efficient Multibit Tree for Large-Scale Chemical Fingerprint Searches

Yasuo Tabei
tabei.y.aa@m.titech.ac.jp,

ERATO Minato Project, Japan Science and Technology Agency, Sapporo, Japan

Abstract

Chemically similar molecules tend to have similar molecular functions. This means molecular functions can be predicted by searching for similar molecules in databases. Thus, similarity searches of chemical compounds are an important research topic in chemoinformatics [1]. The number of available molecules has been constantly increasing. For example, more than 30 million molecules are now stored in the National Center for Biotechnology Information (NCBI) PubChem database. Since the size of the whole chemical space has roughly been estimated to be 10^{60} molecules [1], it will certainly continue to grow after this.

Molecules are typically represented by bit strings that summarize information on molecules. Such representations are called molecular fingerprints. Using binary variables, fingerprints enable us to record the presence or absence of particular functional groups or combinatorial features. Jaccard similarity, also called Tanimoto similarity, has commonly been used in chemoinformatics [2] to measure the similarity between fingerprints.

Several approaches have been proposed to improve similarity searches of fingerprints. As far as we know, most of them have employed the bounds for Jaccard similarity to reduce the number of calculations of similarity [3], [4], [5], [6]. Although these methods proposed tight upper bounds, they did not present an efficient data structure to use these bounds, which resulted in limited scalability with respect to database size.

Thomas et al. [7], [8] solved the efficiency problem by introducing the *pointer-based multibit tree* (MT) which is an efficient tree-based data structure built on the upper bound of [5], [6]. By clustering the database and then building a binary tree by recursively splitting fingerprints for each cluster with the upper bound information, MT enables fast searches of a given query by pruning out useless portions of the search space. Although MT has both theoretical [9] and empirical grounds [8], [9] for efficiency, it has a serious issue with the memory bottleneck caused by the pointers required for its tree-structured implementation, resulting in limited scalability of memory. Moreover, an original fingerprint database needs to be stored in memory to filter out false positives. Since the number of available molecules is ever increasing, developing algorithms using smaller amounts of memory currently remains a challenge.

We present a *succinct multibit tree* (SMT) in this paper, which is a novel compact representation of a multibit tree and fingerprint databases. SMT leverages the idea behind succinct data structures [10] that achieve space-efficient representations of data structures while preserving the property of efficient operations. While the multibit tree and fingerprint databases themselves are represented by succinct data structures, their auxiliaries, e.g. labels, are not always small. In such cases, memory usage is dominated by the auxiliaries to the data structures. To prevent this, we present a novel succinct variable-length array in which elements are represented by bit strings of different lengths. The main difficulty in designing such an array is how to retain the addressability to any element in $O(1)$ time. We overcame this difficulty with the assistance of fast operations in a succinct data structure.

We applied our SMT successfully in experiments to 30 million chemical compounds from PubChem and demonstrated significantly better memory efficiency than that with the pointer-based representation while performing fast similarity searches. In fact, our SMT reduced memory by a factor of 10 compared to the pointer-based approach, resulting in a requirement for only 2 GB of memory for a database managing 30 million fingerprints. The main drawback with our succinct-based approach could be the increase in search time due to multiple calls of operations. We therefore experimented with trade-offs between memory usage and search speed. Surprisingly, our approach was only a few times slower than MT and remained practical, if about 4 GB of memory was used, which is the memory in commercially available PCs.

Acknowledgment

Y.T is supported by MEXT Kakenhi 20589824, the research fellowship from the Japan Society for the Promotion of Science (JSPS) for Young Scientists.

References

[1] M.J. Keiser, B.L. Roth, B.N. Armbruster, P. Ernsberger, J.J. Irwin, and B.K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, 25(2):197–206, 2007.

[2] A.R. Leach and V.J. Gillet. *An introduction to chemoinformatics*. Kluwer Academic Publishers, The Netherlands, rev. ed., 2007.

[3] S. Swamidass and P. Baldi. Bounds and Algorithms for Exact Searches of Chemical Fingerprints in Linear and Sublinear time. *Journal of Chemical Information and Modeling*, 47:302–317, 2007.

[4] P. Baldi, D.S. Hirschberg, and R.J. Nasr. Speeding Up Chemical Database Searches Using a Proximity Filter Based on the Logical Exclusive-OR. *Journal of Chemical Information and Modeling*, 48:1367–1378, 2008.

[5] R. Nasr, D.S. Hirschberg, and P. Baldi. Hashing Algorithms and Data Structures for Rapid Searches of Fingerprint Vectors. *Journal of Chemical Information and Modeling*, 50:1358–68, 2010.

[6] Z. Aung and S.K. Ng. An Indexing Scheme for Fast and Accurate Chemical Fingerprint Database Searching. In *Scientific and Statistical Database Management*, pages 288–305. Springer, 2010.

[7] K. Thomas, N. Jesper, and P. Christian. A Tree Based Method for the Rapid Screening of Chemical Fingerprints. In *WABI*, pages 194–205, 2009.

[8] K. Thomas, N. Jesper, and P. Christian. A tree-based method for the rapid screening of chemical fingerprints. *Algorithms for Molecular Biology*, 5, 2010.

[9] R. Nasr, T. Kristensen, and P. Baldi. Tree and hashing data structures to speed up chemical searches: Analysis and experiments. *Molecular Informatics*, 30:791–800, 2011.

[10] G. Jacobson. Space-efficient Static Trees and Graphs. In *Proceedings of the 30th Annual Symposium of Foundations of Computer Science*, pages 549–554, 1989.

Gene Expression Analysis in Polyploid Species using Next-Generation Sequencer

Jun Sese*, Satoru Akama*, Megumi Yamada†,
Rie Shimizu-Inatsugi‡, and Kentaro K. Shimizu‡
sesejun@cs.titech.ac.jp, akama@ss.cs.titech.ac.jp,
yamada.megumi@sel.is.ocha.ac.jp,
{rie.inatsugi,kentaro.shimizu}@ieu.uzh.ch

* Department of Computer Science, Tokyo Institute of Technology
† Department of Computer Science, Ochanomizu University
‡ Institute of Evolutionary Biology and Environmental Studies, University of Zurich

Abstract

Recent technological advances in sequence technologies called next-generation sequencers (NGS) provide us with billions of DNA short fragment sequences. Such sequencing power has led to new applications of machine learning, data-mining and algorithm techniques to analyze sequences. One of these applications is large-scale gene expression analysis in non-model organisms.

To date, the most common method used to observe gene expression profiles is the microarray [1]. Microarrays work in the following manner (companies selling microarrays have their own methods): A single microarray chip has more than 40,000 spots, and each spot contains identical DNA oligomers. The oligomer is designed to capture (hybridize) only one target gene. The number of copies of the captured gene in each spot is proportional to the strength of the fluorescent signal attached to the tail of the DNA sequence, and the signal level can be measured by using a digital camera and image analysis software.

To design the probes, a genome or entire gene sequences are essential to guarantee that the probe will only hybridize to the target gene. When you use a microarray chip with a nontarget species, mishybridization may cause errors in the interpretation of the expression of the target genes. Despite the problems, much research on large-scale gene expression in non-model organisms has been conducted by using a microarray chip with DNA from a closely related species. The number of species with a reported genome sequence is approximately 3,000[2] although Mora *et al.*[3] predicted that 8.7 million species live on earth. Therefore, the microarray technique has been applied to such a very few organisms with genome sequence information.

With the appearance of NGSs, the microarray is now becoming replaced by the RNA-seq technique[4]. The RNA-seq reads more than 10 millions fragments of mRNAs and then searches for the original positions of these fragments by aligning the sequences against the genome sequence. By counting the number of reads on each gene, we can determine the expression levels of genes because the number of expressed genes is proportional to the number of aligned reads against the genome.

You may think the method can only be applied to model species because the method uses the genome sequence to find the original position of the reads. However, by using a closely related-species genome sequence and search algorithms with tolerance for sequence mismatches[5], we can relate the reads with genes. (Most of the gene sequences between closely related species are similar, tough not identical.) When the target species is not closely related to a genome-sequenced species, the RNA-seq assembly[6] may be used, which reconstructs RNA sequences from RNA-seq reads. The technique is similar to genome assembly but is complicated by the alternative splicing of mRNAs and different expression levels between genes. These techniques allow us to observe gene expression levels in non-model organisms.

Such changes in gene expression observation may allow us to analyze fundamental biological phenomena by comparing between closely related species: the formation of new functions. We are currently conducting research to address this issue by analyzing gene expression changes following whole genome duplication in allopolyploid species

of plants. Allopolyploids are polyploids with chromosomes derived from different species, and many allopolyploids have phenotypes of both parental species.

In this presentation, we will introduce a genome analysis of allopolyploid species using NGS, and then report current results on the measurement of gene expression profiles in allopolyploids using RNA-seq. In measuring gene expression levels, high sequence similarity between homologous genes originating from different parental species prevents us from directly applying the RNA-seq analysis with alignment to a closely related species genome and the RNA assembly. To overcome this limitation, we developed a new method that uses genome sequence of a model species closely related to the target species, and aligned the RNA-seq sequences to this genome. The alignment result contains many mutations, but the mutation might be categorized into three different classes: difference between parental species, difference between target species and used genome, and sequence errors. We are able to classify mutations based on the mutation patterns of the alignment result, and then determines the origin species of each read. To avoid ambiguity of the determination of the origin species coming from sequence errors and errors in alignment position, we used the maximum likelihood estimation method. By counting the number of sequences of each parental species on each gene and using our method, we are able to successfully quantify expression levels of genes in allopolyploid species.

References

[1] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467—470. 1995.

[2] GOLD: Genomes Online Database `http://www.genomesonline.org/`

[3] C. Mora, D.P. Tittensor, S. Adl, A.G.B. Simpson, and B. Worm. How many species are there on Earth and in the ocean? *PLoS Biol* 9: e1001127. 2011.

[4] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57—63. 2009.

[5] M. David, M. Dzamba, D. Lister, L. Ilie, and M. Brudno. SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics* 27: 1011—1012. 2011.

[6] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644—652. 2011.

An Empirical Bayesian Framework for Mutation Detection from Cancer Genome Sequencing Data

Yuichi Shiraishi[*], Yusuke Sato[†], Yusuke Okuno[†], Kenichi Yoshida[†],
Yasunobu Nagata[†], Masashi Sanda[†], Seishi Ogawa[†] and Satoru Miyano[*]
yshira@hgc.jp

[*] Laboratory of DNA Information Analysis, Human Genome Center,
Institute of Medical Science, The University of Tokyo
[†] Cancer Genomics Project, Graduate School of Medicine, The University of Tokyo

Abstract

Recent advances in high-throughput sequencing technologies have enabled a comprehensive dissection of cancer genomes, clarifying large numbers of somatic mutations in a wide variety of cancer types ([1]). Currently, a number of methods have been proposed for mutation calling based on large amount of sequencing data, which is accomplished in most cases by statistically evaluating the difference in allele frequencies of possible SNVs between tumors and matched normal samples ([2], [3]). However, accurate detection of mutations remains still challenging in those situations suffering from low sequencing depths or low tumor burdens. Due to sequencing errors and mapping artifacts accurate somatic mutation calling requires an appropriate statistical analysis. Here, we propose a novel method for detecting somatic mutations to accommodate such problems. Unlike previous methods, it discriminates somatic mutations from sequencing errors based on an empirical Bayesian model for sequencing errors, where model parameters are estimated through the sequencing data from normal samples. Our method not only outperforms existing methods in calling mutations having high-moderate allele frequencies but also achieves highly accurate calling for those mutations having low allele frequencies or harboring in minor fractions of tumor components, allowing for deciphering fine substructures within a tumor specimen.

Acknowledgment

This work is supported by Systems Cancer Project of the Ministry of Education, Culture, Sports, Science and Technology. The super-computing resource was provided by Human Genome Center, Institute of Medical Science, University of Tokyo.

References

[1] K. Yoshida, M. Sanada, Y. Shiraishi, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, 478:64–294, October, 2011.

[2] D. E. Larson, C. C. Harris, K. Chen, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28:311–317, Feburary, 2012.

[3] D. C. Koboldt, Q. Zhang, D. E. . Larson, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22:568–576, March, 2012.

Learning from Treatment History to Predict Response to Anti-HIV Therapy

Hiroto Saigo
saigo@bio.kyutech.ac.jp

Department of Bioinformatics and Bioscience, Kyushu Institute of Technology

Abstract

Infections with the human immunodeficiency virus type 1 (HIV-1) are treated with combinations of drugs. HIV responds to the treatment by developing resistance mutations. For ensuring an effective treatment the genome of the viral target proteins is sequenced and inspected for resistance mutations. For predicting response to a combination therapy, currently available computer-based methods rely on the genotype of the virus and the composition of the regimen as input. However, they do not take full advantage of the knowledge about the order of and the response to previously prescribed regimens. The proposed machine learning system is trained for exploiting such knowledge utilizing the recent advance in frequent sequence mining and support vector machines. When applied to predicting the latest treatment outcome of 3,759 treatment-experienced patients from the EuResist data, prediction accuracy was boosted from 77% to 81%, which constitutes a statistically significant improvement. The major discovery obtained by analyzing the discriminative treatment records is that the information on the composition of a regimen coupled with their short time treatment outcome is as valuable as the composition of a regimen coupled with genotype of the virus. We show that the decision made by our machine learning system is based on clinically relevant rules; such as predicting negative for patients who have already experienced a treatment failure by using standard regimen, or predicting negative for patients who have received regimens which are nowadays known to be ineffective or rather toxic.

Acknowledgment

The authors would like to thank Koji Tsuda and Yoshihiro Yamanishi for fruitful discussion.

References

[1] H. Saigo, A. Altmann, J. Bogojeska, F. Mueller, S. Nowozin and T. Lengauer Learnig from past treatments and their outcome improves prediction of in vivo response to anti-HIV therapy Statistical Applications in Genetics and Molecular Biology, 10(1), 2011.

Machine Learning Methods to Analyze and Infer Drug-Target Interaction Networks

Yoshihiro Yamanishi*
yamanishi@bioreg.kyushu-u.ac.jp

* Division of System Cohort, Multi-scale Research Center for Medical Science,
Medical Institute of Bioregulation, Kyushu University

Abstract

Most drugs are small chemical compounds which interfere with the behavior of their target proteins. Therefore, the genome-wide detection of drug-target interactions or more generally compound-protein interactions from heterogeneous biological data is a key area in chemogenomics and genomic drug discovery toward identification of new drug leads and therapeutic targets for known diseases such as cancers.

In this research, we investigate the correlation between the chemical space of drugs (e.g., chemical structures, fragments), the genomic space of genes or proteins (e.g., sequences, domains, functional sites), and the pharmacological space of phenotypic effects (e.g., efficacy, side-effects, adverse drug reactions) in terms of the topology of drug-target interaction networks. We then develop a new method to predict unknown drug-target interactions from chemical, genomic, and pharmacological data on a large scale. The prediction is performed based on the state-of-art machine learning technology, assuming that drug molecules with similar chemical structures and similar phenotypic effects are likely to interact with similar target proteins. The originality of this research lies in the formalization of the drug-target interaction inference as a supervised learning problem for a bipartite graph, the lack of need for 3D structure information of the target proteins, and in the integration of chemogenomic approach and pharmacogenomic approach in a unified framework. Our comprehensively predicted drug-target interaction networks enable us to suggest many potential drug-target interactions and compound-protein interactions.

References

[1] Pauwels, E., Stoven, V., and Yamanishi, Y., Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*, 12:169, 2011.

[2] Yamanishi, Y., Pauwels, E., Saigo, H. and Stoven, V., Extracting sets of chemical substructures and protein domains governing drug-target interactions. *Journal of Chemical Information and Modeling*, Vol.51(5), pp.1183-1194, 2011.

[3] Lodhi, H. and Yamanishi, Y. *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques*. IGI Global, 2010.

[4] Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, Vol.26, pp.i246-i254, 2010.

[5] Bleakley, K. and Yamanishi, Y., Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, Vol.25, pp.2397-2403, 2009.

[6] Yamanishi, Y., Supervised bipartite graph inference. *Advances in Neural Information Processing Systems 21 (Koller, D., Schuurmans, D., Bengio, Y. and Bottou, L. eds.)*, pp.1841-1848, MIT Press, Cambridge, MA, 2009.

[7] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M., Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, Vol.24, pp.i232-i240, 2008.

Factorized Asymptotic Bayesian Inference for Learning Latent Variable Models

Ryohei Fujimaki*, Satoshi Morinaga†
rfujimaki@sv.nec-labs.com, morinaga@cw.jp.nec.com

* Department of Media Analytics, NEC Laboratories America
† Knowledge Discovery Research Laboratories, NEC

Abstract

In this talk, we present recently-developed factorized asymptotic Bayesian inference (FAB) for learning latent variable models [1], [2] (e.g., mixture models and hidden Markov models.) An interesting and important advantage of FAB over the other methods like the expectation maximization (EM) algorithm or variational Bayesian methods is that it automatically identifies an appropriate model (e.g., the number of clusters in mixture models) as well as model parameters, without any hand-determined parameters.

From an algorithmic viewpoint, an update procedure of FAB is almost the same as that of the EM algorithm. In fact, only difference is that FAB has an exponentiated regularization in each E step as described below. This difference is small in terms of the update procedure, but makes essential difference from the other methods.

	EM Algorithm for mixture models		FAB for mixture models
E step :	$q^t_{nc} \propto \alpha^{t-1}_c p(\boldsymbol{x}_n \vert \phi^{t-1}_c)$	E step :	$q^t_{nc} \propto \alpha^{t-1}_c p(\boldsymbol{x}_n \vert \phi^{t-1}_c) \exp(-D_c/2\alpha^{(t-1)}_c N)$
M step :	$\alpha^t_c \propto \sum_{n=1}^N q^t_{nc}$ $\phi^t_c = \arg\max_\phi \sum_{n=1}^N q^t_{nc} \log p(\boldsymbol{x}_n \vert \phi^{t-1}_c)$	M step :	$\alpha^t_c \propto \sum_{n=1}^N q^t_{nc}$ $\phi^t_c = \arg\max_\phi \sum_{n=1}^N q^t_{nc} \log p(\boldsymbol{x}_n \vert \phi^{t-1}_c)$

From theoretical viewpoints, FAB has several desirable properties: 1) asymptotic consistency with the marginal log-likelihood, 2) automatic component selection on the basis of an intrinsic shrinkage mechanism, 3) capability of optimizing component types as well as the number of components, and 4) parameter identifiability in latent variable models. While this talk focuses on intuitive explanation and algorithmic aspects of FAB, we will briefly introduce these theoretical properties.

References

[1] R. Fujimaki and S. Morinaga Factorized Asymptotic Bayesian Inference for Mixture Modeling. In *Proceedings of Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012

[2] R. Fujimaki and K. Hanashi Factorized Asymptotic Bayesian Hidden Markov Models. In *Proceedings of 29th International Conference on Machine Learning (ICML)*, 2012

Distances and Kernels on Discrete Structures

Marco Cuturi
mcuturi@i.kyoto-u.ac.jp

Graduate School of Informatics, Kyoto University

Abstract

Distances and positive definite kernels lie at the core of many machine learning algorithms. When comparing vectors, these two concepts form well-matched pairs that are almost interchangeable: trivial operations such as changing signs, adding renormalization factors, taking logarithms or exponentials are usually sufficient to recover one from the other (e.g. Euclidean distances & Laplace kernels) [1]. However, when comparing discrete structures, this harmonious symmetry falls apart. The culprit lies in the introduction of combinatorial optimization to compute distances (e.g. edit distances for strings, time series or trees; minimum cost matching distances for sets of points; transportation distances for histograms etc.). Simple counterexamples show that such considerations – finding a minimal cost matching or a maximal alignment to compare two objects – tend to destroy any hope of recovering a positive definite kernel from such distances. We present a review of several results in the recent literature that have overcome this limitation, including recent [2], [3], [4] and unpublished work. We provide a unified framework for these approaches by highlighting the fact that they all rely on generating functions to achieve positive definiteness.

References

[1] C. Berg, J.P.R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Number 100 in Graduate Texts in Mathematics. Springer Verlag, 1984.

[2] J.-P. Vert, H. Saigo, and T. Akutsu. Local alignment kernels for protein sequences. In Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, 2004.

[3] M. Cuturi, J.-P. Vert, Ø. Birkenes, and T. Matsui. A kernel for time series based on global alignments. In *Proceedings of the Intern. Conference on Acoustics, Speech and Signal Processing*, volume II, pages 413 – 416, 2007.

[4] K. Shin, M. Cuturi, and T. Kuboyama. Mapping kernels for trees. *Proc. of ICML 2011*, 2011.

Efficient Enumeration of Bounded-Size Subtrees in a Tree and Its Application to Tree Mining with Proximity Constraint

Kunihiro Wasa*, Yusaku Kaneta*[1], Takeaki Uno†, and Hiroki Arimura*
{wasa, y-kaneta, arim}@ist.hokudai.ac.jp, uno@nii.jp

* Graduate School of Information Science and Technology, Hokkaido University
† National Institute of Informatics

Abstract

By emergence of massive structured data, there have been increasing demands for efficient methods that discovers interesting patterns or regularity hidden in collections of structured data ([1], [2]). *Frequent tree and graph mining* ([2]) is one of the well studied framework of such semi-structured data mining, where the task is to find all subgraphs that appear in input labeled trees at least specified times by preserving their labels and topology. There are extensive researches on efficient frequent tree and graph mining algorithms for the classes of ordered trees ([3], [4]), unordered trees and free trees ([5], [2]), and graphs ([1], [6], [4]), and applications to knowledge discovery from real structured data (See, e.g. [7], [8]). On the other hand, the topology preserving constraint of frequent tree mining is often considered too strict to apply tree mining algorithms to real structured data in, e.g., biology or natural language processing, since the structure of a motif is sometimes not well preserved in the field.

To overcome this difficulty, recent researches in pattern matching and mining ([9]) have considered the introduction of *proximity constraint* instead of topology constraint, where a sort of proximity should be retained, while the shape can change. For instance, in the *graph motif problem* is a proximity version of tree matching problem, originally introduced by (Lacroix, Fernandes, and Sagot [9]). In the problem, given a multiset of labels P as a pattern and a large node labeled tree T as a text, we are requested to find a connected subgraph H of T whose multiset of labels is identical to P, regardless of the shape of H. Similar extension of tree mining can be easily imagined.

By the motivation of such proximity pattern discovery, we study a special case of the k-subtree enumeration problem for trees, originally introduced by (Ferreira, Grossi, and Rizzi [10]) for general input graphs, where an input graph is a tree of n nodes, and we are requested to find all distinct k-subtrees (subtrees consisting of exactly k nodes) contained in an input tree. We present the first *constant delay* enumeration algorithm that lists all k-subtrees of an input tree in $O(1)$ worst-case time per subtree. Combining label matching and frequency counting with this enumeration algorithm, we obtain an efficient frequent tree pattern mining under the proximity constraint, where a pattern is a multiset of labels whose occurrences are connected in an input tree. Furthermore, we will discuss application of our algorithm to the graph motif problem for trees, too.

In Fig. 1, we show an example of the search space generated by the algorithm for all k-subtrees from S_1 to S_{19} in an input tree T_1. Our constant-delay algorithm enumerates all k-subtree in an input tree by starting from a special initial tree (S_1 in the figure) whose nodes are consecutively numbered from 1 to k, and by applying two expansion rules, called expansions of type I and type II, respectively. Unlike the *rightmost expansion technique* for ordered tree mining (e.g., [3], [4]) , traverse from a k-subtree P (called a *parent*) to another k-subtree Q (called a *child*) is done by deleting a leaf from P and adding a *border node* (any child of some node in P that does not belong to P, yet) to P to keep the size of P constant. By expansion rule of type I (thin solid arrows in the figure), the algorithm enumerates a group of the k-subtrees that share the same node as their roots. For instance, a group $\{S_2, S_3\}$ share node 2 as their roots, while another group $\{S_1, S_4, \ldots, S_{19}\}$ share node 1 as their roots. By expansion rule of type II (thick dashed arrows in the figure), the algorithm traverses all different groups from one to other by generating k-subtrees of special form, called *serial trees*, whose nodes are consecutively numbered. In the figure, S_1, S_2, S_9 and S_{10} are such serial trees as the root of groups. By careful management of candidates for the nodes to add and to delete using their characterization, we can finally implement the algorithm using constant

[1] Presently, he is working for Rakuten Research, Tokyo, Japan.

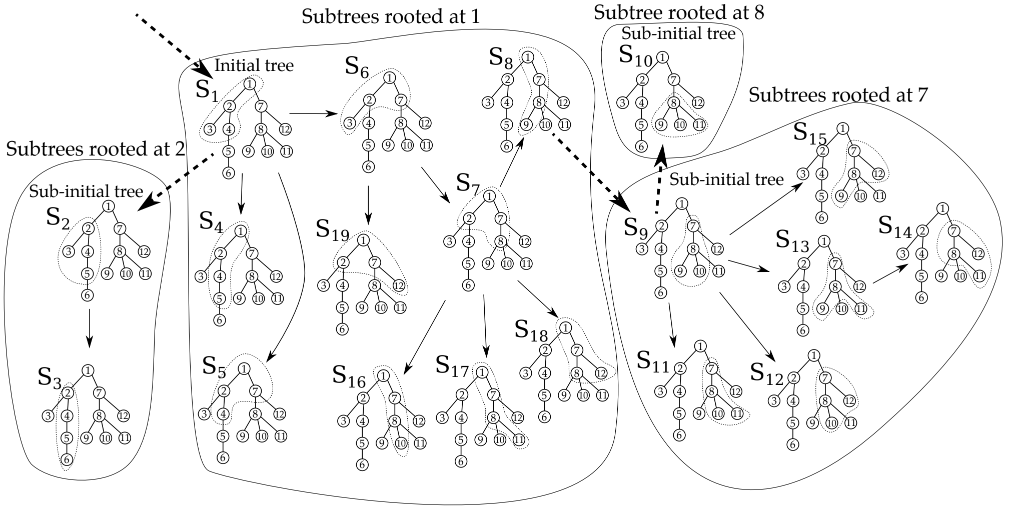

Figure 1. The search space of all nineteen k-subtrees of an input tree T_1 of size $n = 11$, where the size of a subtree is $k = 4$. In this figure, each set of nodes in a tree surrounded by a dotted circle indicates a k-subtree. Solid arrows and dashed arrows indicate expansion rules of type I and type II, respectively.

update time. In summary, we presented an efficient algorithm for enumerating k-subtrees in a tree with application to tree mining with proximity constraint. This result improves on Ferreira et al's $O(k)$ amortized time algorithm in the case that an input is a tree.

Acknowledgment

The authors would like to thank Shin-ichi Nakano, Takuya Kida, Shin-ichi Minato, and Thomas Zeugmann for their valuable comments. This research was partly supported by MEXT Grant-in-Aid for Scientific Research (A), 20240014 (FY2008–2011), 20240021 (FY2012–2015), and MEXT/JSPS Global COE Program (FY2007–2011).

References

[1] Takashi Washio and Hiroshi Motoda. State of the art of graph-based data mining. *SIGKDD Explorations*, 5(1):59–68, 2003.

[2] Yun Chi, Richard R. Muntz, Siegfried Nijssen, and Joost N. Kok. Frequent subtree mining - an overview. *Fundam. Inform.*, 66(1-2):161–198, 2005.

[3] Tatsuya Asai, Kenji Abe, Shinji Kawasoe, Hiroki Arimura, Hiroshi Sakamoto, and Setsuo Arikawa. Efficient substructure discovery from large semi-structured data. In *Proc. the 2nd SIAM Int'l Conf. on Data Mining (SDM'02)*. SIAM, 2002.

[4] Mohammed Javeed Zaki. Efficiently mining frequent trees in a forest. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pages 71–80. ACM, 2002.

[5] Tatsuya Asai, Hiroki Arimura, Takeaki Uno, and Shin-Ichi Nakano. Discovering frequent substructures in large unordered trees. In *Proc. Discovery Science 2003 (DS'03)*, volume 2843 of *LNCS*, pages 47–61. Springer, 2003.

[6] Xifeng Yan and Jiawei Han. gSpan: Graph-based substructure pattern mining. In *Proc. the 2002 IEEE International Conference on Data Mining (ICDM'02)*, pages 721–724. IEEE, 2002.

[7] Koji Tsuda and Taku Kudo. Clustering graphs by weighted substructure mining. In *ICML 2006*, pages 953–960, 2006.

[8] Satoshi Morinaga, Hiroki Arimura, Takahiro Ikeda, Yosuke Sakao, and Susumu Akamine. Key semantics extraction by dependency tree mining. In *Proc. KDD'05*, pages 666–671. ACM, 2004.

[9] Vincent Lacroix, Cristina G. Fernandes, and Marie-France Sagot. Motif search in graphs: Application to metabolic networks. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 3:360–368, 2006.

[10] Rui Ferreira, Roberto Grossi, and Romeo Rizzi. Output-sensitive listing of bounded-size trees in undirected graphs. In *Proc. ESA'11*, volume 6942 of *LNCS*, pages 275–286, 2011.

Fast Similarity Search with Succinct Trees

Koji Tsuda*
koji.tsuda@aist.go.jp

* Computational Biology Research Center,
National Institute of Advanced Industrial Science and Technology
† JST ERATO Minato Project

Abstract

In the last 10-15 years there has been a great increase of interest in space-efficient (succinct) data structuresthat are compressed up to the information theoretic lower bound. Compared to pointer-based naive data structures, the memory usage can be smaller up to 20-30 fold. I briefly present basics of succinct data structures and our recent work of indexing 25 million chemical graphs for similarity search in memory.

Simultaneous Aligning and Folding of RNA Sequences via Dual Decomposition

Kengo Sato[*], Yuki Kato[†], Tatsuya Akutsu[‡], Kiyoshi Asai[§¶] and Yasubumi Sakakibara[*]
satoken@bio.keio.ac.jp, ykato@is.naist.jp,
takutsu@kuicr.kyoto-u.ac.jp, asai@k.u-tokyo.ac.jp, yasu@bio.keio.ac.jp

[*] Department of Biosciences and Informatics, Keio University
[†] Graduate School of Information Science, Nara Institute of Science and Technology
[‡] Institute for Chemical Research, Kyoto University
[§] Graduate School of Frontier Sciences, University of Tokyo
[¶] National Institute of Advanced Industrial Science and Technology

Abstract

Many functional RNAs form secondary structures that are related to their functions such as gene regulation and maturation of mRNAs, rRNAs and tRNAs. Since experimental determination of RNA secondary structures is very expensive and time-consuming, computational prediction of RNA secondary structures is frequently utilized. It is well known that the accuracy of RNA secondary structure prediction from single sequences is limited, thus the comparative approach that predicts common secondary structures from aligned sequences is a better choice if homologous sequences with reliable alignments are available. However, we require correct secondary structure information for producing reliable alignments of RNA sequences. This is a chicken-and-egg problem. To tackle this problem, several algorithms that simultaneously align and fold RNA sequences, also known as Sankoff-based algorithms [1], have been proposed (e.g. Murlet [2], RAF [3]). However, Sankoff-based algorithms are still computationally expensive although various techniques for reducing the computation time have been developed.

We develop a novel algorithm that simultaneously aligns and folds RNA sequences. First, on the basis of maximizing expected accuracy (MEA) principle, we design an objective function for pairwise structural alignments as the expectation of the sum of the number of correctly predicted base-pairs in common secondary structures and the number of correctly aligned columns in pairwise alignments. Then, in order to maximize the objective function under several constraints for consistent pairwise structural alignments, we employ the dual decomposition technique, which decomposes the pairwise structural alignment problem into two secondary structure prediction problems and one pairwise (non-structural) alignment problem. The algorithm maintains the consistency of pairwise structural alignments by imposing penalties on inconsistent base pairs and alignment columns, and updates them iteratively by performing the Nussinov-style secondary structure prediction and the Needleman-Wunsch-style pairwise alignment with the penalized scoring function. We can easily apply the progressive alignment technique to extend the algorithm into multiple alignments. Furthermore, we can apply several standard techniques for multiple alignments including the iterative refinement [4] and the probabilistic consistency transformation (PCT) [5] into our algorithm. Due to the dual decomposition technique, the secondary structure model and the pairwise alignment model are highly independent of each other. This allows us to easily take pseudoknots into account by replacing the secondary structure model with the IPknot model [6].

In order to confirm the effectiveness of our algorithm, we conducted the computational experiments on the Murlet dataset [2] that includes only pseudoknot-free structures, and the Rfam-PK dataset [6] that includes pseudoknotted structures. We evaluated the accuracy of predicted structural alignments through the sum-of-pairs score (SPS), the structure conservation index (SCI) for predicted alignments, and the sensitivity (SEN), the positive predictive value (PPV) and the Matthews correlation coefficient (MCC) for predicted base pairs. The total computation time (TIME) was measured on a Linux workstation with Intel Xeon E5450 (3.0GHz). Table 1 shows the result on the Murlet dataset, comparing our algorithm, called `DAFS`, with state-of-the-art algorithms including `CentroidAlign` [7], `RAF` [3], `LARA`[8] and `LocARNA` [9]. This result indicates that `DAFS` achieved the best in terms of MCC out of all the aligners, and is much faster than `RAF`, one of the most accurate competitors. Table 2 shows the result on the Rfam-PK dataset. Note that there is no existing practical structural aligner that can consider pseudoknots

Table 1. The result on the Murlet dataset.

	SPS	SCI	SEN	PPV	MCC	TIME (s)
DAFS	0.75	0.46	0.67	0.77	**0.71**	416
CentroidAlign	**0.78**	0.48	0.62	0.80	0.69	169
RAF	0.75	0.46	**0.68**	0.75	**0.71**	4274
LARA	0.75	0.50	0.62	0.78	0.68	5361
LocARNA	0.71	**0.61**	0.64	0.76	0.69	14540
ProbConsRNA	0.76	0.37	0.56	**0.84**	0.66	88

Table 2. The result on the Rfam-PK dataset.

	SPS	SCI	SEN	PPV	MCC	TIME (s)
DAFS						
(IPknot decoding)	0.89	**0.78**	**0.62**	**0.67**	**0.64**	242
(Nussinov decoding)	**0.90**	**0.78**	0.60	0.66	0.62	44
RNASampler	0.81	0.73	0.59	0.65	0.61	2783

except for RNASampler [10]. The result clearly shows the advantage of DAFS with IPknot decoding that can simultaneously align and fold RNA sequences with pseudoknots. Furthermore, the comparison between Nussinov decoding and IPknot decoding indicates that IPknot is successfully integrated into DAFS, meaning that DAFS is flexible and extensible due to the dual decomposition.

Acknowledgment

The authors wish to thank Dr. Michiaki Hamada and our colleagues from the RNA Informatics Team at the Computational Biology Research Center (CBRC) for fruitful discussions. The super-computing resource was provided by Human Genome Center, Institute of Medical Science, University of Tokyo.

References

[1] D. Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appli. Math.*, 45(5):810–825, 1985.

[2] H. Kiryu, Y. Tabei, T. Kin, and K. Asai. Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, 23(13):1588–1598, July 2007.

[3] C. B. Do, C.-S. Foo, and S. Batzoglou. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, 24(13):i68–76, July 2008.

[4] M. P. Berger and P. J. Munson. A novel randomized iterative strategy for aligning multiple protein sequences. *Computer applications in the biosciences : CABIOS*, 7(4):479–484, Oct. 1991.

[5] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research*, 15(2):330–340, Feb. 2005.

[6] K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):i85–i93, July 2011.

[7] M. Hamada, K. Sato, H. Kiryu, T. Mituyama, and K. Asai. CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics*, 25(24):3236–3243, Dec. 2009.

[8] M. Bauer, G. W. Klau, and K. Reinert. Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinform.*, 8:271, 2007.

[9] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3:e65, Apr 2007.

[10] X. Xu, Y. Ji, and G. D. Stormo. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics*, 23:1883–1891, Aug 2007.

Frequent Pattern Mining for Families of Dispersed Repeats in DNA Sequences

Atsuyoshi Nakamura*, Ichigaku Takigawa†, Hisashi Tosaka*, Mineichi Kudo*
and Hiroshi Mamitsuka‡
{atsu,t_hisashi,mine}@main.ist.hokudai-u.ac.jp,
takigawa@cris.hokudai.ac.jp, mami@kuicr.kyoto-u.ac.jp

* Graduate School of Information Science and Technology, Hokkaido University
† Creative Research Institution, Hokkaido University
‡ Institute for Chemical Research, Kyoto University

Abstract

We develop a method of systematically finding families of dispersed repeats in long sequences using frequent pattern mining. In DNA sequences, such families are known as retrotransposons, which are genetic elements that can amplify themselves. Thought many retrotransposons have been already known, no systematic search has been done yet.

In the framework of frequent pattern mining, we try a systematic search for families of dispersed repeats in long sequences. In order to catch family members only as occurrences of an approximate pattern, we adopt the counting method that counts *locally optimal* occurrences alone [1], [2]. In the mining, only substrings with maximal alignment score are counted as occurrences of a pattern string. For checking local optimality of occurrences, alignments of both forward and backward directions are necessary, which force the enumeration algorithms to spend $O(n^4)$ time or $O(n^3)$ space, where n is the length of a given string.

For faster solution, we propose a k-gap constrained version of the problem and develop an $O(n^3)$-time and $O(n^2)$-space algorithm and its *candidate-based* version. By limiting the number of gaps to at most k, local optimality for both directions can be checked simultaneously, which reduces the space that is necessary for keeping candidate occurrences to be checked for the opposite direction. The candidate-based version calculates alignment scores for candidate neighborhood only, which is fast when the number of candidates for long patterns is small.

We applied the 1-gap constrained candidate-based version to each of 24 human chromosomes (Homo_sapiens. GRCh37.64.dna.chromosome.x.fa for $x = 1, 2, ..., 22$, X and Y). We enumerated all the maximal approximate frequent patterns whose length is at least 100 and whose number of occurrences is at least 100 per length 50,000,000 from each chromosome. After removing all the non-approximately-maximal patterns, only 25 patterns were obtained, and 23 of them are families of dispersed repeats. One of them is similar to the most common SINEs(Short INterspersed Elements) called Alu family. Most of the others are parts of the most common LINE(Long INterspersed Elements) called LINE-1.

Acknowledgment

This work was partially supported by the Collaborative Research Program of Institute for Chemical Research, Kyoto University (grant 2012-55).

References

[1] B. W. Erickson and P. H. Sellers. Recognition of patterns in genetic sequences, In David Sankoff and Joseph B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules : The Theory and Practice of Sequence Comparison*, Reading, MA, Addison-Wesley, pp. 55–91 (1983).

[2] A. Nakamura, H. Tosaka and M. Kudo. Mining Approximate Patterns with Frequent Locally Optimal Occurrences. *Hokkaido University Division of Computer Science TCS Technical Report Series A* (http://www-alg.ist.hokudai.ac.jp/tra.html), TCS-TR-A-10-41 (2010).