Random Forest Proximity Clustering of Sequence

Sheng Liu, Qing Song

Abstract—**Clustering biological sequences into functional groups such as protein families, motif families is one of the goals of sequence analysis. One important factor in the cluster analysis of sequences is similarity measure between two sequences. The performance of existing similarity measure methods such as edit distance based methods is limited especially in twilight zone. In this paper, we introduce random forest proximity measure which has been successful in some other fields like tissue classification, DNA microarray data analysis. The random forest proximity measure obtains better specificity on some datasets showing its suitability on such sequence data. These results were comparable with those obtained from alignment score, distance based methods.**

Index Terms— **Random Forests, Clustering, Biological Sequence.**

I. Introduction

The growing biological sequence data provide a wealth of information. A number of computational, statistical and machine learning methods for the analysis of sequence have been developed. There are two main methods for similarity measure [1,2,3]: alignment-based methods and alignments-free methods. Alignment based method such as BLAST [4], edit distance [5] based methods is widely used in database search. As alignment-free method is still in its early stage, much should be studied to get a clearer view of them comparing with alignment-based methods. Similarity measures, such as, single linkage method [6], Euclidean distance [7], Mahalanobis distance [8], Markov chain models, Kullback-Leibler discrepancy [9], information content methods, and so on. In this paper, we present a method based on both alignment and non-alignment methods to get the similarity measure. An alignment score is usually calculated by summing the scores for each aligned position and the scores for gaps, so it does not consider the dependency in the sequence that may be a key factor making two sequences differ from each other. Clustering based on alignment scores thus can be improved by introducing algorithms that consider dependency between sequence positions.

A random forest is a classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees [10]. It has many properties such as: it will not overfit the data, can deal with data with high dimension (large amount of variables), can detect variable importance from data and can detect outliers. Ref. [11] introduced random forest proximity for clustering. Due to its special similarity measure, groups obtained in this way may be different to the traditional Euclidean metrics, thus offer another view of the data. One important issue is: it considers dependency between sequence positions (variables). As a consequence, it may offer better clustering performance if there are dependency between positions in given data. The random forest proximity has been successful in DNA microarray data[11], tumor marker data [12,13] and protein-protein interaction data [14], etc. Few work that computes proximity between biological sequences has been done. In this paper, we first give a brief introduction on random forest proximity, sequence representation follows and then test the random forest proximity measure on three datasets. The results show that random forest proximity is comparable with alignment score methods and Hierarchical Clustering methods. In some cases, it is even more favorable.

Sheng Liu and Qing Song are with the School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore.

II. Methods and Results

A. Random Forest Proximity

By sampling the same size of sequences as the original set of sequences with the same distribution of each position (variable), the difference between the two sets of sequences are that positions in sampled set of sequences are independent to each other, while positions in the original set of sequences may be dependent [15]. Considering the original set of sequences as class 1 and the synthetic set of sequences as class 2, this can be checked by a two class classification on the two sets of data. In classification, the two sets of sequences are sampled so that two-thirds of the sequences are used as training data, one-third of the cases are left out of the sample to get a unbiased estimate of the error(testing data). If the classification error is low, as the only difference in the two sets of sequences is position dependence, the position dependence plays an important role in the discrimination of the two sets. At the same time, if two sequences in the training data are in the same terminal node of a tree in the classification process, we can increase their proximity by one. At the end, the normalized proximities are used as input in clustering as (dis)similarity measure. The details can be seen [11].

B. *Representation of The Sequences*

So far one of the most effective way in sequence representation is to use multiple alignment. Alignment score, profiles, Position Specific Scoring Matrix(PSSM) are all based on it. As the sequence length variance is not too large in our datasets, it is better first align all the sequences in the dataset. Each aligned sequence is of uniform length with gaps filled. From algorithm point of view, the letters in each position are category data. Random forest also can deal with them properly.

C. *The Random Forest Proximity Clustering Framework*

One of the major parts of clustering algorithms is to determine the dissimilarity measure. Random forest proximity matrix is generated from the classification of the sequence sets. We can then map it to lower dimension space via multidimensional scaling [16]. After that deterministic annealing (DA) algorithm is applied to approach global optimum. Fig. 1 illustrates the framework. R [17] package randomForest [18] are used. R codes of DA are given upon request to authors.

D. *Experiments on E. coli K12 DNA binding sites*

Dataset is from [19], E. Coli DNA binding matrices. Only a binding protein has 15 or above corresponding DNA binding sites are chosen for analysis, leading to 154 binding sites corresponding to crp, purR, lexA, tyrR, argR, metJ, phoB 7 proteins. DNA sequences are represented by frequency of occurrence over background frequency. We also do alignment using Clustral W [20] on the dataset. The alignment scores obtained are regarded as similarity measure. The frequency represented dataset is also clustered by Hierarchical Clustering(R package hclust) directly. Random forest proximity calculation is based on the code from [21] while using DA for clustering to approach global optimum. The results are as in Fig.2-Fig. 4 and Table 1. In Fig. 2, the grid on the diagonal is well-proportioned, while the other two figures are less. Showing random forest proximity method has better discrimination in the dataset. With different similarity method within Hierarchical Clustering, the results are also different (data not shown here), but still more poor than random forest proximity method. Accuracy is measured by (true positive + true negative)/Total size, Sensitivity is measured by true positive/(true positive + false negative), specificity is measured by true negative/(true negative + false positive). For each binding protein, a true positive is correctly clustered as that binding protein, a false positive is falsely clustered as that binding protein, a true negative is correctly clustered as not being that binding protein and finally a false negative is falsely clustered as not that binding protein. Multidimensional scaling of the proximity data to the two dimension plot (Fig. 5) showing 4 cluster is already well separated.

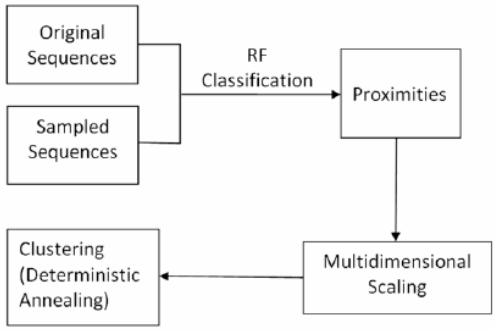

Fig.1. Framework of Random Forest Proximity Clustering.

TABLE 1
ACCURACY OF THREE METHODS ON E. COLI K12 DNA BINDING SITES

Random Forest Proximity	*Alignment score*	Hierarchical Clustering
87.4%	68.2%	47.1%

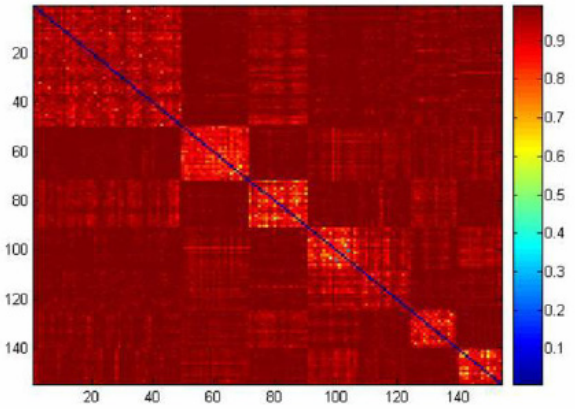

Fig. 2. Plot of similarity matrix generated from random forests clustering.

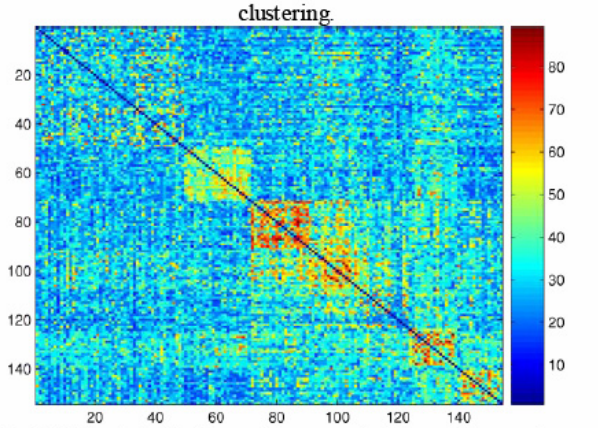

Fig 3. Plot of similarity matrix generated from Alignment score

E. *Experiment on FoldAlign data*

Dataset is from [22] on FoldAlign test set for global clustering. The results are shown as in Table 2. Sequences in dataset are aligned with Clustral W. Filling the gap positions in the alignments with 0, we get an equal length aligned data

for each set. Although the cluster size vary considerably, it still has some robust to cluster bias. Random Forest Proximity method is a bit better on average than alignment score methods.

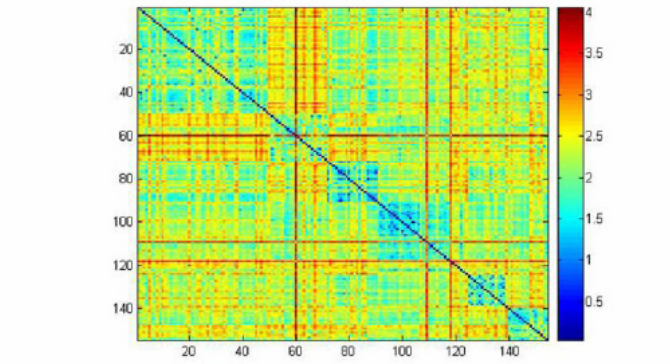

Fig. 4. Plot of similarity matrix generated from Hierarchical Clustering.

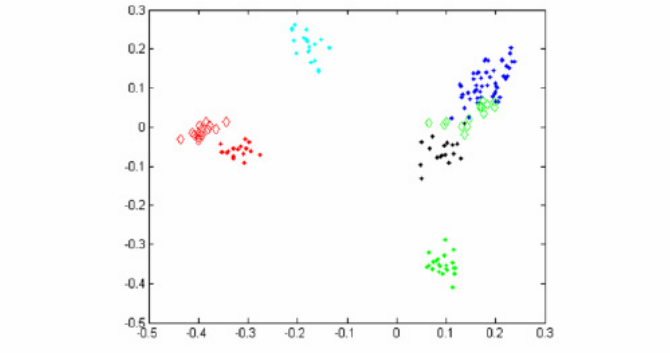

Fig. 5. Plot of cluster of sequences via Multidimensional scaling.

III. DISCUSSION AND CONCLUSION

On some sequence data we obtain better performance, better specificity. The reason for this maybe: it captures dependency difference between two classes. On the contrary, through classification of the two sets of sequences, we can have an idea if there is position dependency among sequences. The position dependency possibly comes from sequences' function. For example, two positions in a binding site may need contact with transcription factor at the same time to let the transcription factor function. These two positions then are related (both should be conserved) showing dependency. Random forest has intrinsic variable selection to select important positions that relate to the sequence functions such as binding positions, RNA structure domains, which is to be further analyzed. It provides another promising choice when selecting suitable algorithms for the clustering analysis of sequence data.

TABLE 2
PERFORMANCE OF TWO METHODS ON FOLDALIGN DATA

Methods	Accuracy	Sensitivity	Specificity
RF Proximity +	**99.19**	**93.00**	**99.55**
Alignment Score	98.99	90.89	99.46
AS manhattan	98.99	90.89	99.47

ACKNOWLEDGMENT

We thank reviewers to give valuable comments and suggestions.

REFERENCES

[1] T.F. Smith, M.S. Waterman, "Identification of Common Molecular Subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.

[2] S. Vinga, and J. Almeida, "Alignment-free sequence comparison - a review," *Bioinformatics*, vol. 19, pp. 513–523, 2003.

[3] T. Pham and J. Zuegg, "A probabilistic measure for alignment-free sequence comparison," *Bioinformatics*, vol. 20, pp. 3455–3461, 2004.

[4] S.F. Altschu, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic local alignment search tool," *J. Mol. Biol*., vol. 215, pp.403–410, 1990.

[5] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Doklady Akademii Nauk SSSR*, vol. 163, pp. 845–848, 1965.

[6] V. Veeramachaneni, and W. Makalowski, "Visualizing sequence similarity of protein families," *Genome research*, vol. 14, pp. 1160–1169, 2004.

[7] B.E. Blaisdell, "A measure of the similarity of sets of sequences not requiring sequence alignment," *Proc. Natl Acad. Sci. USA*, vol. 83, pp. 5155–5159, 1986.

[8] T.J. Wu, J.P. Burke, and D. B. Davison, "A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words," *Biometrics*, vol. 53, pp. 1432–1439, 1997.

[9] T.J. Wu, Y.C. Hsieh, and L.A. Li, "Statistical measures of DNA dissimilarity under Markov chain models of base composition," *Biometrics*, vol. 57, 441–448, 2001.

[10] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[11] L. Breiman, and A. Cutler, "Random Forests Manual v4.0, " Technical report, UC Berkeley, available online at ftp://ftp.stat.berkeley.edu/pub/users/breiman/Using random forests v4.0.pdf. , 2003.

[12] T. Shi, D. Seligson, A.S. Belldegrun, A. Palotie, and S. Horvath, "Tumor Classification by Tissue Microarray Profiling: Random Forest Clustering Applied to Renal Cell Carcinoma," *Modern Pathology*, vol. 18, pp. 547–557, 2005.

[13] D.B. Seligson, S. Horvath, T. Shi, H. Yu, S. Tze, M. Grunstein, and S.K. Kurdistani, "Global Histone Modification Patterns Predict Risk of Prostate Cancer Recurrence," *Nature*, vol. 435, 1262–1266, 2005.

[14] Y. Qi, J. Klein-Seetharaman, Z. Bar-Joseph, "Random forest similarity for protein-protein interaction prediction from multiple sources," *Pac Symp Biocomput*., pp. 531–542, 2005.

[15] Y. Barash, G. Elidan, N. Friedman, T. Kaplan, "Modeling dependencies in protein-DNA binding sites," *In Proceedings of the seventh annual international conference on Computational molecular biology*, ACM Press New York, NY, USA. pp. 28-37, 2003.

[16] M.F. Cox, and M.A.A. Cox, "Multidimensional Scaling," Chapman and Hall , 2001.

[17] R. Ihaka, and R. Gentleman, "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, vol. 5 pp. 299–314, 1996.

[18] A. Liaw, and M. Wiener, "Classification and Regression by randomForest", R News: The Newsletter of the R Project vol. 2, pp. 18–22, 2002.

[19] K. Robison, A.M. McGuire, G.M. Church, "A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K12 genome," *Journal of Molecular Biology*, vol. 284, pp. 241–254, 1998.

[20] J.D. Thompson, D.G. Higgins, and T.J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, pp. 4673–4680, 1994.

[21] T. Shi, and S. Horvath, "Unsupervised Learning with Random Forest Predictors," *Journal of Computational and Graphical Statistics*, vol. 15, pp. 118–138, 2006.

[22] E. Torarinsson, J.H. Havgaard and J. Gorodkin, "Multiple structural alignment and clustering of RNA sequences," *Bioinformatics,* Bioinformatics Advance Access published online on February 25, 2007