Poster H-67

Decomposition of DNA Sequences into Hidden Components: Applications to Human Genome's Promoter Recognition

Authors:
Yasuo Matsuyama (*Waseda University*)
Kenji Onuki (*Waseda University*)
Youichi Ito (*Waseda University*)
Yugo Ishihara (*Waseda University*)
Kouhei Kawasaki (*Waseda University*)
Takayuki Hasegawa (*Waseda University*)

Short Abstract: Decomposition of discrete-symbol biosequences into hidden components is discussed. Presented methods are based upon the principal component analysis and the independent component analysis. The decomposition procedure is described and tested on human genome's promoters and other specific sites to show better recognition scores than traditional non-decomposing methods.

Long Abstract:

1. Addressed Problem and Significance: Discrete-symbol biosequences composed of DNA, RNA or amino acids have patterns emerging from sets of neighboring elements. Since each pattern is related to specific functions, finding these clusters or motifs has become a strong interest of bioinformatics. For such a problem, simple string matching methods usually fail. This is because a single segment in a pattern may have polymorphic functions. Besides, the function may vary depending upon environmental and/or statistical backgrounds. Therefore, it is meritorious to find components which are hidden in discrete-symbol biosequences. In this presentation, the principal component analysis (PCA) and the independent component analysis (ICA) are used for the decomposition. Fig. 1 illustrates the concept of the decomposition. There are n components or soft sub-segments obtained by the decomposition. Each segment is described by a weighted summation of these components. The weights are utilized for the motif recognition. This method is realized as a set of software and is experimented on the recognition of the human genome's promoters and other functional sites. Better performances than existing non-decomposing methods are obtained.

2. Decomposition and Recognition Methods: The decomposition method is based on PCA and ICA [1]. Fig. 2 illustrates the procedures of the learning and recognition phases for the human genome's promoter recognition. The step for the decision on the CpG-Related/Non-Related [2] exists depending on the GC richness [3]. Following this decision, the training segments are converted to a set of numerals using the position dependent frequency. Then, the data are preprocessed by PCA. The resulting data set is further processed by ICA to find the hidden components. By this ICA learning, the decomposition matrix Wtr and the superposition coefficient matrix Ytr are obtained. These matrices are stored and used in the recognition phase to compute the feasibility score of the promoter estimation. It is important to note here that the recognition step of Fig. 2 is hierarchical by two levels. Since the method does not assume any known TSS, the estimation of this site is necessary. This phase also use the ICA decomposition.

3. Experiments: The PCA and ICA hidden component decomposition methods are tested on human genome's promoters, TSS and poly-A signals. Here, the scores of the R/N decision, the initiator decision and the TATA box decision are computed and used to generate the total score for the promoter recognition. The current status is the following. On the promoter, the performance is (PREC, SPEC, SENS) = (0.632, 0.573, 0.824) which is better than (not reported, 0.535, 0.793) of [2]. For the TSS, the performance is (PREC, SPEC, SENS) = (0.691, 0.725, 0.656). For the poly-A signal, the precision was 0.987. Thus, the presented recognition method based on the hidden component decomposition of the discrete-symbol sequence was found to be effective.
4. Discussions and Concluding Remarks: The main purpose of this presentation was to show the applicability and the utility of the decomposition of discrete-symbol biosequences into hidden components. To show this idea's properness, the recognition of the human genome's promoters was addressed. The resulting scores were better than existing non-decomposing methods. Thus, the idea of the discrete-symbol decomposition to soft sub-patterns was shown to be viable. There are further sophistications for the increase of the recognition rates. Such ideas include (1) semi-supervised methods, and (2) the extension to multiple clusters. Finally, we point out the following as further possibilities: (a) Since PCA and ICA are common tools apart from the physical or chemical entities of the target data, the discrete-symbol version with the symbol softening can be a good tool for the front or rear engine to existing bioinformatics tools. (b) Experiments on amino acid sequences are worthy to try.
[1] Matsuyama Y, Katsumata N, Kawamura R (2003). Independent component analysis minimizing convex divergence. Lecture Notes in Computer Science 2714: 7-34.
[2] Davuluri R, Grosse I, Zhang MQ (2001). Computational identification of promoters and first exons in the human genome. Nature Genetics 29: 412-417.
[3] Hannehalli S, Levy S (2001). Promoter prediction in the human genome. Bioinformatics 17, Suppl. 1: S90-96.

Fig. 1 Hidden component decomposition (equivalent to 140 words) is here.

Fig. 2 ICA learning for promoter recognition (equivalent to 160 words) is here.