Genomic Data Mining for Species Identification Using Principal Component Analysis

Shreyas Sen,[1] Seetharam Narasimhan,[1] Amit Konar,[1] Uday K. Chakraborty[2]

[1]Department of Electronics and Telecommunication Engineering,
Jadavpur University, Kolkata-32, India
[2]Dept. of Math & Computer Science, University of Missouri, St. Louis, MO 63121, USA, uday@cs.umsl.edu

Abstract

This paper aims at designing a scheme for automatic identification of a species from its genome sequence. A set of 64 three-tuple keywords is first generated using the four types of bases: A, T, C and G. These keywords are searched on N randomly sampled genome sequences, each of a given length (10,000 elements), and the frequency count for each of the $4^3 = 64$ keywords is obtained. Principal component analysis is then employed on the frequency counts for N sampled instances. The principal component analysis yields a unique feature descriptor for identifying the species from its genome sequence. The variance of the descriptors for a given genome sequence being negligible, the proposed scheme finds applications in automatic species identification.

Keywords: Genome sequence, principal component analysis, data mining, species identification, feature descriptor diagram

1. Introduction

Genomic data mining and knowledge extraction is an important problem in bioinformatics. Identification of a species from its genomic database is a challenging task. This paper explores a new approach to extract genomic features of a species from its genome sequence.

Extensive work on local and global alignment of DNA sequences has already been done by a number of researchers. The Smith-Waterman algorithm [1, 2] for local alignment and the Needleman-Wunsch algorithm [3] for global alignment of DNA-sequences are two well-known algorithms for matching genome databases. These works, though extremely valuable, have their limitations. The demerits include the use of complicated matrix algebra and dynamic programming, and the results of sequence matching are not free from pre-calculated threshold values. It is to be noted that none of the above-mentioned methods can be directly employed to identify the species from the structural signature of the genomes.

Rapid advances in automated DNA sequencing technology [4] have generated the need for statistical summarization of large volumes of sequence data so that efficient and effective statistical analysis can be carried out. The popular sequence alignment algorithms and techniques for estimating homologies [5] and mismatches among DNA sequences that are used for comparing sequences of relatively small sizes are not applicable to sequences with sizes varying between a few thousand base pairs to a few hundred thousand base pairs. Even for comparison of small sequences, the standard alignment and matching algorithms are known to be time consuming. There is a dearth of rapid and parsimonious procedures that may be somewhat approximate in nature yet useful in producing quick and significant results. The present paper is an attempt to fill this void. The idea is to make the analysis of large DNA sequences easier by statistically summarizing the data using dimensional reduction, while capturing some of the fundamental structural information contained in the sequence data.

To the best of the authors' knowledge, identifying a species from its genomic data is an open problem. The novelty of the work reported in this paper is as follows. First, the paper takes into account frequency counts of 64 three-lettered primitive DNA attributes in randomly selected samples of the genome sequences of different species (e.g., the bacterium Escherichia coli (E. coli) [6], Drosophila melanogaster [7], Saccharomyces cerevisiae (yeast) [8] and Homo sapiens (human beings)). Second, to reduce the data dimension of extracted features (here, frequency counts), principal component analysis (PCA) [9] is employed on the randomly selected samples of

genome sequence. Third, the variance of the extracted feature vectors being extremely small for any randomly selected input sequence, the accuracy of the results in identifying the species is very high.

The paper is divided into 5 sections. In section 2, we outline the proposed scheme for generating genomic features of a species from its genome sequence. Section 3 provides the detailed steps of using PCA in the proposed application. Simulation results are presented in section 4. Conclusions are drawn in section 5.

2. The Proposed Scheme for Feature Extraction from Genomic Data

The method of feature extraction from genomic data is outlined below:

1. Construct the possible keywords of length 3 using the literals A (Adenine), C (Cytosine), T (Thymine) and G (Guanine). The number of such keywords is $4^3 = 64$. The 3-tuples are 'AAA', 'AAC', 'AAT', 'AAG' ,'ACA',..., 'GGT', 'GGG' .

2. Randomly select N samples, each of length 10,000 bases, from a given genome sequence.

3. Determine the frequency count of each keyword in each string. To illustrate what we mean by frequency count, let us take an example. Consider a small portion of the sequence like …AATCG…. It contributes a count of 1 to the frequency of occurrence of each of the three keywords AAT, ATC and TCG. Similarly, for the substring …TTTTT…, we get a count of 3 for the frequency of the keyword TTT. Proceeding similarly for a large sample sequence of 10,000 bases, we get frequencies of all the 64 keywords in the form of a frequency count vector of dimension (1×64).

4. Apply PCA on the frequency count data for N samples to reduce dimension and get the most significant data.

The first three steps in the above algorithm are self-explanatory. The fourth step needs some explanation. In the next section we expand the sub-steps in the 4th step. Here we only state its significance.

It is important to note that the frequency counts of 64 three-element keywords in a 10,000 element string of genome sequence are more or less invariant with respect to the random sampling of the genome sequence. Naturally, our main emphasis of study was to determine whether the small difference in the counts of a given keyword in N samples is statistically significant. PCA provides a solution to this problem. First, the dimension of (N×64) is reduced by PCA to (1×64). Second, the (minor) disparity in the feature gets eliminated by PCA. Since PCA is a well-known tool for data reduction without loss of accuracy, we believe that our results on feature extraction from the genome database are also free from loss of accuracy.

3. How PCA is Used in the Present Context

The methodology of employing PCA to the given problem is outlined below:

Input: A set of N vectors (1×64) representing the frequency counts of 64 three-tuple keywords.

Output: A minimal feature descriptor vector sufficient to describe the problem without any significant loss in data.

1. Normalization: Let the i^{th} (1×64) input vector be denoted by

$$\boldsymbol{a_i} = [a_{i1} \;\; a_{i2} \;\; \ldots\ldots \;\; a_{i64}]$$

To get the vector normalized we use the following transformation:

$$a_{ik} \leftarrow \frac{a_{ik}}{\sum_{j=1}^{64} a_{ij}}$$

2. Mean adjusted data: To get the data adjusted around zero mean, we use the formula:

$$a_{ik} \leftarrow a_{ik} - \overline{a_i} \qquad \forall i,k$$

where $\overline{a_i}$ =mean of the i^{th} vector

$$= \frac{1}{64} \sum_{j=1}^{64} a_{ij}$$

The matrix (N×64) so obtained is called the *Data Adjust*:

$$\textit{Data Adjust} = \begin{pmatrix} a_{11} & \cdots & a_{164} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{N64} \end{pmatrix}$$

3. Evaluation of the covariance matrix: The covariance between any two vectors a_i and a_j is obtained by the following formula:

$$\mathrm{cov}(a_i, a_j) = c_{ij} = \frac{\sum_{k=1}^{64} \left(a_{ik} - \overline{a_i}\right)\left(a_{jk} - \overline{a_j}\right)}{(n-1)}$$

Covariance matrix C for the N different (1×64) vectors is represented as follows:

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \ddots & \vdots \\ c_{N1} & \cdots & c_{NN} \end{pmatrix}$$

where C is an N×N matrix.

4. Eigenvalue evaluation: From the roots of the equation | C -- λI | = 0, the eigenvalues of the

covariance matrix C are obtained. There would be N eigenvalues of matrix C, and corresponding to each eigenvalue there would be eigenvectors each of dimension N×1.

5. Principal component evaluation: The eigenvalues are not the same. In fact, it turns out that the eigenvector corresponding to the *highest* eigenvalue λ_{large} is the *principal component* (N×1) of the data set. Therefore

$$\textit{Principal Component} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix}$$

where

$$\lambda_{large} > \lambda_i \qquad 1 \le i \le N$$

6. Projection of data adjust along the principal component: To get the *feature descriptor*, the following formula is applied:

$$\textit{Feature Descriptor} = \textit{Principal Component}^{\mathbf{T}} \times \textit{Data Adjust}$$

where Principal Component^T (1×N) is the transpose of the *Principal Component* vector.

Thus we get a *Feature Descriptor* vector of dimension 1×64 corresponding to N samples of the genome sequence database of the particular species.

7. Computing the mean feature descriptor: We calculate M such *feature descriptors* from different random samples and then calculate the mean of these vectors and also the variance vector (both 1×64).

4. Results: Geometric Representation of Feature Descriptor

The *feature descriptor diagrams* for different species are described here. We could represent the *feature descriptors* using bar diagrams, pie-charts or any other standard representation. However, using the polar plot we get figures that are compact yet distinct representations of the *mean feature descriptor*.

As mentioned earlier, the *mean feature descriptor* is a 1×64 vector. So to construct these diagrams 360^0 is divided into 64 equal parts, corresponding to 64 keywords. Plotting it in polar (r,ɵ) co-ordinates with r as the values of the *mean feature descriptor* vector and ɵ as these angles we get the *feature descriptor diagrams.*

The *feature descriptor diagrams* are distinctly different from species to species. So we can readily detect new species and identify known species by comparing their *feature descriptor diagrams*.

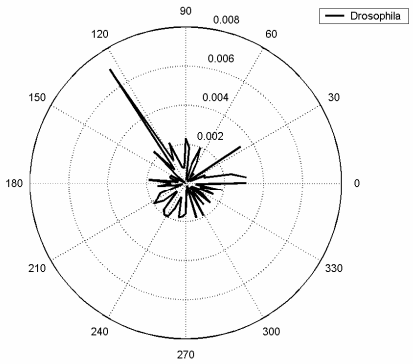

Fig 1: Feature Descriptor Diagram for Drosophila.

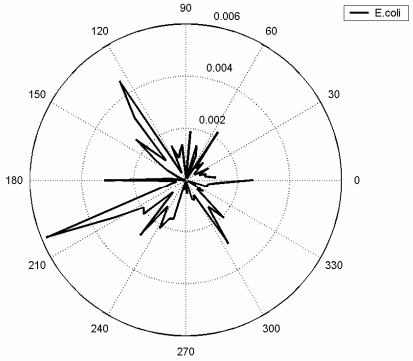

Fig 2: Feature Descriptor Diagram for E. coli.

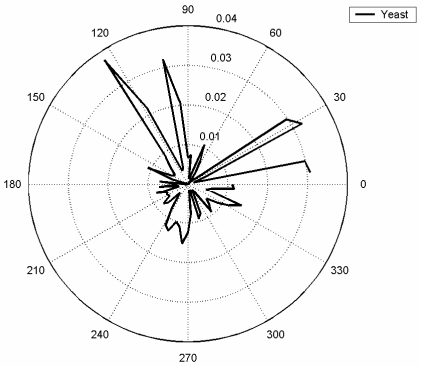

Fig 3: Feature Descriptor Diagram for Yeast.

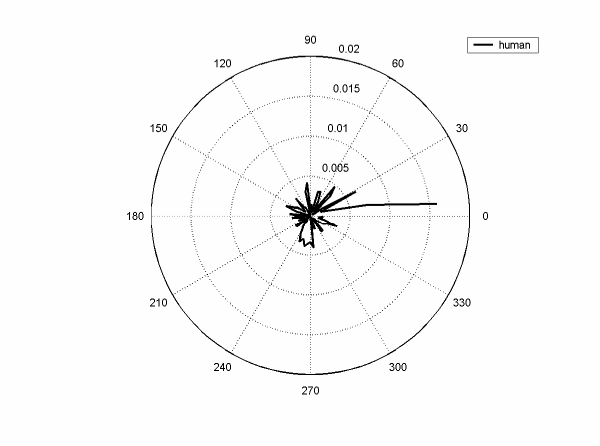

Fig 4: Feature Descriptor Diagram for human chromosome rp11-433k2.

If we closely observe Fig 1 which contains the diagrammatic representation of the mean feature descriptor vector for Drosophila we can see some distinct peaks with a prominent one at around 125˚. In contrast, Fig 2 drawn for E. coli has its highest peaks near 125˚ and 200˚ and smaller ones around 140˚, 180˚ and 300˚. Similar distinctions are clearly visible from the other diagrams as well.

5. Conclusions

To our knowledge, this is the first work of its kind to extract information from complete genome sequences and to distinguish between species by feature descriptor diagrams.

The micro array of DNA sequences for a species follows a repetitive pattern of nucleotide bases. This paper emphasizes the above statement by representing the micro array sequence by a specialized *feature descriptor vector*. The mapping of the DNA arrays to feature descriptors need not be unique. In fact, any type of nonlinear mapping that compresses the large DNA array to a vector of very small dimension could be employed to correlate the structural topology of *feature descriptor* with a given species.

In this process we have used PCA to reduce the large dimensions of genome sequence data without loss of accuracy. If only the frequency count is plotted then we do get some difference from species to species but it is not enough to distinguish between them.

This is where PCA comes in. When PCA is applied to the original data we get enough differences between the *feature descriptor diagrams* of different species that enable us to tell one species from another with the help of these diagrams.

By constructing *feature descriptor diagrams* for several species it has been observed that they are quite different from species to species, indicating that we can certainly associate each figure with each species and claim that the figure represents the species. Moreover when *feature descriptor vectors* for the same species are calculated, they turned out to be nearly identical, with insignificant variance.

There is further scope of using this method for distinguishing between two different chromosomes of the same species.

Acknowledgments

Thanks to three anonymous referees for their comments.

References

[1] Smith, T. F., and Waterman, M. S., "Identification of common molecular subsequences," *J. Mol. Biol.* pp.147, 195-197, 1981.

[2] Waterman, M. S., and Eggert, M., "A new algorithm for best subsequence alignments with applications to tRNA-rRNA,*" J.Mol.Biol.*, pp. 197, 723-728, 1987.

[3] Needleman, S. B., and Wunsch, C. "A general method applicable to the search for similarities in the amino sequence of two proteins." *J. Mol. Biol.* pp. 48,443-453, 1970.

[4] Galperin, M. Y., and Koonin, E. V., Comparative Genome Analysis, In *Bioinformatics- A Practical Guide to the Analysis of Genes and Proteins,* Baxevins, A. D., and Oullette, B. F. F.(Eds.), Wiley-Interscience, New York, 2nd ed., p. 387, 2001.

[5] States, D. J., and Boguski, M. s., Similarity and Homology, In *Sequence analysis primer* Gribskov, M., and Devereux, J. (Eds.), Stockton Press, New York, pp. 92-124, 1991.

[6] Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., et al., "The complete genome sequence of Escherichia coli," Vol. K-12, *Science,* pp. 277, 1453-1462.

[7] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., et al, "The genome sequence of Drosophila melanogaster," *Science* pp. 287, 2185-2195, 2000.

[8] Cherry, J. M., Ball, C., Weng, S., Juvik, G., Schimidt, R., Alder, C., Dunn, B., Dwight, S., Riles, L. et al., "Genetic and Physical maps of Saccharomyces cerevisiae," *Nature* (suppl. 6632) pp. 387, 67-73, 1997.

[9] Smith, L. I., "A tutorial on Principal Components Analysis," 2002.