Visualization and Classification of DNA Sequences Using Pareto Learning Self Organizing Maps Based on Frequency and Correlation Coefficient

Hiroshi Dozono

Department of Advanced Fusion, Saga University,
1 Honjyo Saga 840-8502 Japan
hiro@dna.ec.saga-u.ac.jp

Abstract. Next-generation sequencing techniques produce an enormous amount of sequence data. Analyzing these sequences requires an efficient method that can handle large amounts of data. Self-organizing maps (SOMs), which use the frequencies of N-tuples, can categorize sets of DNA sequences with unsupervised learning. In this study, SOM using correlation coefficients among nucleotides was proposed, and its performance was examined in the experiments through mapping experiments of the genome sequences of several species and classification experiments using Pareto learning SOMs.

1 Introduction

Next-generation sequencing [1] produces large amounts of sequence data that are applied to many areas of genome science. Meta-genome and comparative genome analyses are examples of such applications. Meta-genome analysis uses mixtures of genomes from a group of species for analysis of the composition of species or expressed sequences. Comparative genome analysis uses the sequenced genome data of a group of species to analyze evolutionary relationships or species diversity. Both applications require a global comparison of DNA sequences among species.

Self organizing maps(SOMs)[2] are often used for the global comparison of DNA sequences. SOMs are neural networks that use the architecture of feed-forward networks and train the network with an unsupervised learning method. A set of input vectors is given to the network, and SOM extracts the features of the input vectors on two-dimensional maps according to vector similarity.

The frequencies of N-tuples, which denote the occurrence of each N-tuple for a fixed N, are effective for global comparison, and we proposed an analysis of DNA sequences with an SOM by using the vectors of N-tuple frequencies as input vectors [3]. For large-scale data, the use of these frequencies as feature SOM vectors is effective, and it is also applied to the analysis of IP-packet traffic For large scale data, it is effective to use the frequencies as feature vector of SOM,

T. Villmann et al. (eds.), *Advances in Self-Organizing Maps and Learning Vector Quantization,* Advances in Intelligent Systems and Computing 295,
DOI: 10.1007/978-3-319-07695-9_8,

and it is also applied to the analysis of the traffic of IP-packets [4]. In a previous study [3], the relationships among the genomes of species were visualized on the basis of frequencies of N-tuples. Further research proceeded using this method.

Herein, we propose another preprocessing method on the basis of correlation coefficients (CCs) of the occurrences of each nucleotide in a DNA sequence. All combinations between 2 nucleotides A-A, A-C, A-G, A-T, C-A, ..., T-G, T-T, CCs of the occurrences in the sequences are calculated by shifting 1 of the sequences in 1 to N. For 1 to N shifts, the number of CCs is $4^2 \times N$. CCs are arranged in vectors and used as input vectors for SOM, which determines the global features of the DNA sequences.

Furthermore, we apply Pareto learning SOMs (P-SOMs) [5] to visualize and classify DNA sequences. P-SOMs use a multi-modal vector composed of multiple vectors, including the category vector that denotes the class of the vector for supervised learning. The category vector operates cooperatively with the original input vectors to improve visualization and classification. P-SOMs were examined in the benchmark data set iris [5] and applied to the authentication method for behavior biometrics [6] and IP-packet traffic analysis [4].

2 Pareto Learning Self Organizing Map (P-SOM)

2.1 Pareto Learning SOM for Multi Modal Vector

2.2 P-SOM for Multi-modal Vectors

A multi-modal vector $(\{\boldsymbol{x}_1\}, \{\boldsymbol{x}_2\}, \ldots, \{\boldsymbol{x}_n\})$ is a vector composed of multiple vectors and attributes. For example, keystroke timing and key typing intensity are the features used for authentication with key typing features. In multi-modal vectors, each vector and attribute is described in a different unit and scale, and the availability for the classification may be different. Conventional SOMs can learn multi-modal vectors by using a simply concatenated vector $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$or a concatenated vector with weight values $(w_1\boldsymbol{x}_1, w_2\boldsymbol{x}_2, \ldots, w_n\boldsymbol{x}_n)$ as the input vector. When weight values are excluded, the map is dominated by largely scaled vectors and easily affected by unreliable vectors. A map using weight values depends heavily on these values, making the selection of optimal weight values difficult.

P-SOM makes direct use of a multi-modal vector $\boldsymbol{x} = (\{\boldsymbol{x}_1\}, \{\boldsymbol{x}_2\}, \ldots, \{\boldsymbol{x}_n\})$ as an input vector based on Pareto optimality. For each vector, $\boldsymbol{x}_i$, the objective function is defined as $f_i(\boldsymbol{x}, U^{jk}) = |\boldsymbol{x}_i - \boldsymbol{m}_i^{ij}|$ for unit U^{jk} on the map, where $\boldsymbol{m}^{ij} = (\{\boldsymbol{m}_1^{jk}\}, \{\boldsymbol{m}_2^{jk}\}, \ldots, \{\boldsymbol{m}_n^{jk}\})$ is the vector associated with U^{jk}. The Pareto winner set $P(\boldsymbol{x})$ for an input vector $\boldsymbol{x}$ is the set of the units U^{jk} that are Pareto optimal according to the object functions $f_i(\boldsymbol{x}, U^{jk})$. Thus, P-SOM is a multi-winner SOM and all units in $P(\boldsymbol{x})$ and their neighbors are updated simultaneously.

The algorithm of P-SOM is as follows.

P-SOM Algorithm

1. Initialization of the map
   Initialize the vector $\mathbf{m^{ij}}$ which are assigned to unit U^{ij} on the map using the 1st and 2nd principal components as base vectors of 2-dimensional map.
2. Batch learning phase
   (1) Clear all learning buffer of units U^{ij}.
   (2) For each vector x^i, search for the pareto optimal set of the units $P = \{U_p^{ab}\}$. U_p^{ab} is an element of pareto optimal set P, if for all units $U_{kl} \in P - U_p^{ab}$, existing h such that $e_h^{ab} \leq e_h^{kl}$ where
   $$e_h^{kl} = \left|\mathbf{x_h^i} - \mathbf{m_h^{kl}}\right| \tag{1}$$
   (3) Add x^i to the learning buffer of all units $U_p^{ab} \in P$.
3. Batch update phase
   For each unit U^{ij} update the associated vector $\mathbf{m}^{ij}$ using the weighted average of the vectors recorded in the buffer of U^{ij} and its neighboring units as follows.
   (1)For all vectors x recorded in the buffer of U^{ij} and its neighboring units in distance $d \leq Sn$, calculate weighted sum $\mathbf{S}$ of the updates and the sum of weight values W.
   $$\mathbf{S} = \mathbf{S} + \eta fn(d)(\mathbf{x} - \mathbf{m^{i'j'}}) \tag{2}$$
   $$W = W + fn(d) \tag{3}$$
   where $U^{i'j'}$s are neighbors of U^{ij} including U^{ij} itself, η is learning rate, $fn(d)$ is the neighborhood function which becomes 1 for d=0 and decrease with increment of d.
   (2) Set the vector $\mathbf{m^{ij}} = \mathbf{m^{ij}} + \mathbf{S}/W$.

Repeat 2. and 3. with decreasing the size of neighbors Sn for pre-defined iterations.

Fig.1 shows the differences in the SOM and P-SOM algorithms.

In the update phase, the units in the overlapped neighbors are updated more strongly, and this phase plays an important role in the integration of multimodal vectors. P-SOM is scale free because all vectors in $\boldsymbol{x}$ are handled evenly independently to the scales of $\boldsymbol{x}_i$

P-SOM can integrate any kind of vector. Thus, the category vector c^i can be introduced as an independent vector for each input vector to P-SOM.

$$\mathbf{\acute{x}}^i = (\boldsymbol{x}^{\boldsymbol{i}}, \boldsymbol{c}^{\boldsymbol{i}}) \tag{4}$$

$$c_j^i = \begin{cases} 1 \ \boldsymbol{x}^{\boldsymbol{i}} \in C_j \\ 0 \ otherwise \end{cases} \tag{5}$$

The category vector is also used to search the Pareto winner set, and it attracts the input vectors in the same category that correspond closely on the map with the original input vector $\boldsymbol{x}$. The category of the given test vector $\boldsymbol{x}_t$ is determined as $argmax\{\sum_{U^{ij} \in P(\boldsymbol{x}_t)} c_k^{ij}\}$ where $P(\boldsymbol{x_t})$ is the Pareto optimal set of units for $\boldsymbol{x}_t$.

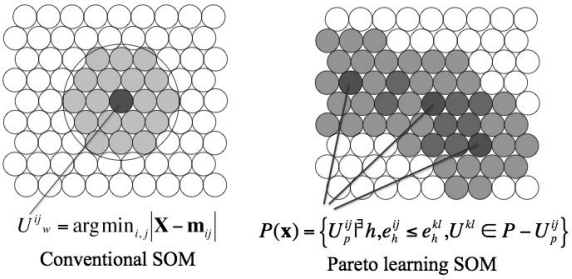

Fig. 1. Differences between the self-organizing map (SOM) and Pareto learning SOM (P-SOM) algorithms

3 Analysis of DNA Sequences Using SOM

This section explains the preprocessing methods for effective extraction of DNA sequencw features.

3.1 Frequency of DNA Sequences

The frequency of N-tuples in DNA sequences is defined as the number of N-tuples in the sequence. Fig. 2 shows an example of the frequency for N = 2. Long

Sequence	AGTAATCTCTAATCT														
Frequency	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG
	1	0	1	2	0	0	0	3	0	0	0	1	1	2	0

Fig. 2. Frequency of the 2-tuple of a DNA sequence

sequences are divided into segments of constant length to enlarge the number of learning vectors. SOMs, which uses the frequency of DNA sequences as the input vector, can reportedly visualize the relationship of the genomes of different species for N = 4 and N = 5 [4]. However, the dimension of the frequency vector becomes 4^N. Thus, for large N values, the size of the input vector becomes very large.

3.2 Correlation Coefficient(CC)s of the Nucleotides in DNA Sequences

A DNA sequence is the sequence of the characters 'A','G','T', and 'C', thus, it is meaningless to calculate the CC directly for the sequence. A DNA sequence is converted to 4 binary sequences that represent the occurrences of every nucleotides 'A', 'G', 'T', and 'C'. For all combinations of the occurrence sequences,

$\rho_{n1,n2}(i)$, which is CC between the first occurrence sequence of nucleotides $n1$ and the sequence that shifts N nucleotides from the second occurrence sequence of nucleotides $n2$ are calculated for i=1 to N. Fig.3 shows the example of the calculation of CCs. These CCs are concatenated in a vector, and used as the input

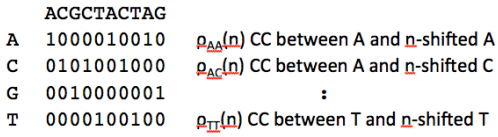

Fig. 3. Correlation Coefficients of DNA sequence

vector for the SOM. Calculating CCs requires the scanning of the sequences 16 times, and has huge computational costs for long sequences. Using the following equation, all CCs of between 2 sequences of nucleotides, $S^1 = s_1^1 s_2^1 \cdots s_L^1$ and $S^2 = s_1^2 s_2^2 \cdots s_L^2$, can be calculated with 1 pass scan.

$$C_1 = \begin{cases} 1 & s_k^1 = n1 \\ 0 & s_i^1 \neq n1 \end{cases} \tag{6}$$

$$C_2 = \begin{cases} 1 & s_i^2 = n2 \\ 0 & s_i^2 \neq n2 \end{cases} \tag{7}$$

$$\sigma_{n1,n2} = \sum_{i=1}^{L} (C_1 - m_{n1})(C_2 - m_{n2}) \tag{8}$$

$$\sigma_{n1,n1} = \sum_{i=1}^{L} (C_1 - m_{n1})^2 \tag{9}$$

$$\sigma_{n1,n2} = \sum_{i=1}^{L} (C_2 - m_{n2})^2 \tag{10}$$

$$\rho_{n1,n2} = \frac{\sigma_{n1,n2}}{\sigma_{n1,n1}\sigma_{n1.n2}} \tag{11}$$

where m_{n1} and m_{n2} are the averages of the occurrence sequences for nucleotides $n1$ and $n2$ respectively.

Compared with the dimensions of the frequency vector, the dimension of the vector is small. It is $16 \times N$ for the concatenated vector of 1 to N shifts.

3.3 Experimental Results

The purpose of applying SOM for the analysis of DNA sequences is visualization. This subsection gives the experimental results of visualization of the relations between DNA sequences based on frequencies and CCs. We used two sets od

DNA sequences. The first set comprised the DNA sequences of 6 species registered to the pathway of amino acid metabolism in the Kyoto Encyclopedia of Genes and Genomes database. These species are colored as shown in Table 1.

Table 1. Species used in the experiments

Genome name	Description	Color
hsa	homo sapience	red
cfa	dog	blue
mmu	mouse	green
dme	fruit fly	yellow
eco	E-Coli	magenta
osa	rice	cyan

Table 2. Pathways used in the experiments

Pathway name	Color
amino acid metabolism	red
cell growth and death	blue
metabolism of complex carbohydrates	green
metabolism of complex lipids	yellow
nucleotide metabolism	magenta
transration	cyan
transcription	white

The second set comprised the DNA sequences of 6 pathways of homo sapience. Gene sequences registered to multiple pathways were removed from the set. In this paper, The pathways are colored as shown in Table 2.

In both sets, the sequences which are longer than 1000 were segmented to the sequences with a length of 1000. The total number of the segments was 7148 for the species set, and 1135 for the pathway set.

The parameters of SOM was given as follows.

- map size: 128×64
- learning rate: from 0.8 to 0.1
- update method: batch update
- neighborhood function: gausian function
- iteration of learning: 50

Fig.4 shows the map of frequencies of 4-tuples. he length of the vector is $4^4 = 256$. Each color dot on the map represents the fragment of the sequence colored

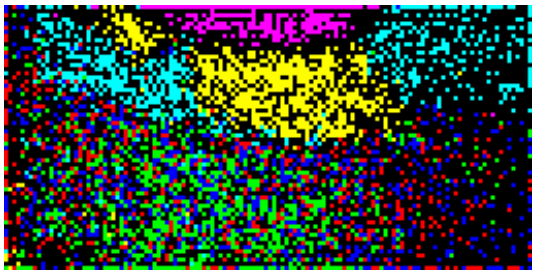

Fig. 4. Map of the frequencies of 4-tuples in the DNA sequences of 6 species

as shown in Table.1. Sequences of dme, eco and osa were clustered separately. Sequences of hsa, cfa and mmu were loosely clustered because they are mammals. Fig.5 shows the map of CC of 1 to 4 shifts. The length of the vector is $16 \times 4 = 64$. The topologies of these maps are similar, and the clarity of the clusters is almost

Fig. 5. Map of CC of 1 to 4 shifts in the DNA sequences of 6 species

the same. When the number of shifts and length of tuples is decreased, as shown in Fig.6 to Fig.9, CCs show better clustering results than those of frequencies.

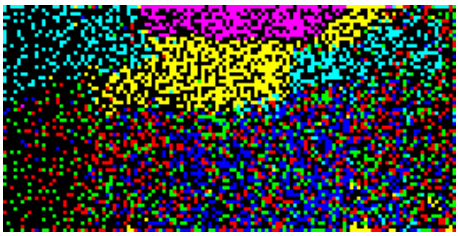

Fig. 6. Map of the frequencies of 3-tuples L=64

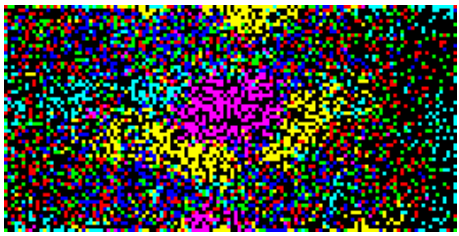

Fig. 7. Map of the frequencies of 2-tuples L=16

Considering the length of the vector(L), CCs represented the features of DNA sequences more effectively than the frequencies of N-tuples did.

Fig.10 and Fig.11 show the maps of the frequencies of 4-tuples and CCs of 1 to 4 shifts using the pathway set respectively.

When the pathway set was used, the sequences were not clustered clearly. However, each color showed the shading in the specific area on the map, which was considered loosely clustered.

As an additional experiment, Fig.12 shows the maps of CCs using the input data of 7 different virus genomes. Some virus genomes are fragmented in some regions, however they are clustered as the species set.

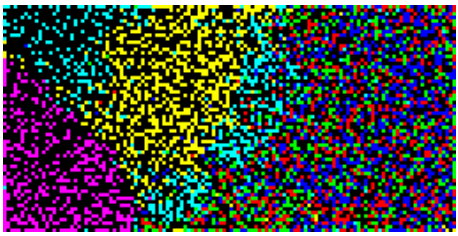

Fig. 8. Map of the CCs of 1 and 2 shifts L=32

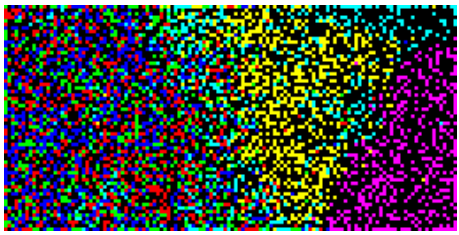

Fig. 9. Map of the CC of 1 shift L=16

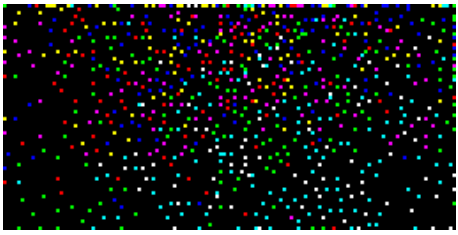

Fig. 10. Map of the frequencies of 4-tuples of the pathway set

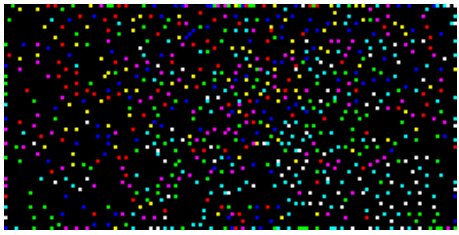

Fig. 11. Map of the CCs of 1 to 4 shifts of the pathway set

4 Analysis of DNA Sequences Using Pareto Learning SOM(P-SOM)

We analyzed the DNA sequences using P-SOM. P-SOM can learn input vectors both in unsupervised learning mode without using category vectors for learning and in supervised learning mode with using category vectors. In supervised learning mode, category vectors cooperate with the original input vectors to organize the map. A vector of 16 CCs for each shift is used as an element of multi-modal input vectors to the P-SOM.

Fig.13 and Fig.14 show the maps of the CCs of 1 to 4 shifts using the species set and the pathway set as input vectors. The maps are torus maps. In Fig.13, the species are clustered, as seen in the results of the conventional SOM, and the mammals are clustered more strongly than those in the conventional SOM because of the supervised learning feature of the P-SOM. In Fig.14, the pathways are also more clearly clustered than those of the conventional SOM.

For the classification experiment, a randomly selected 70 % of the sequences were used for learning, and 30 % of the sequences were used for the test. CCs and frequencies of N-tuples were used as input vectors, and the experiments using conventional SOM were conducted for comparison. Table 3 shows the results for the species set. In this table, CC-N denotes the CC of 1 to N shifts, and F-N

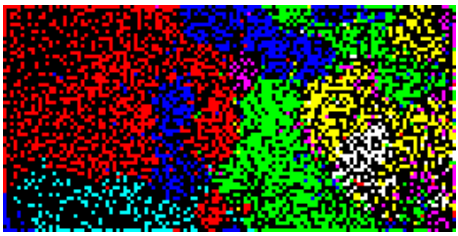

Fig. 12. Map of CC of 1 to 4 shifts of 7 virus genome

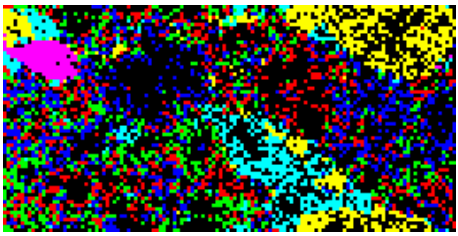

Fig. 13. Map of the CCs of 1 to 4 shifts by using the species set as the input vectors for P-SOM

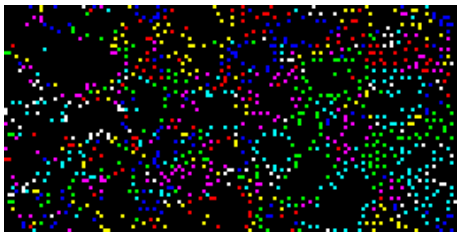

Fig. 14. Map of the CC of 1 to 4 shifts using the pathway set as the input vectors for P-SOM

Table 3. Rates of successful classification of the species set

Input vector	CC-2	CC-4	CC-2	CC-4	F-4	F-4
Length	32	64	32	64	256	256
Method	P-SOM	P-SOM	SOM	SOM	P-SOM	SOM
Learned sequences	0.832	0.831	0.920	0.916	0.980	0.915
Test Sequences	0.609	0.643	0.593	0.599	0.624	0.629

denotes the frequency of N-tuples. For the learned sequences, the P-SOM using frequency as the input vector performed best, and for the test sequences, the P-SOM using CCs performed best. Table 4 shows the rates of successful classification for each species. As expected, the rates for mammals are poor because they were loosely clustered on the map. The sequences from cfa(dog) may be miss classified to hsa and mmu. The accuracy seems to be low as the classifier, because the coding regions of mammals include common genes. For the virus genome set, the accuracies for learned sequences and test sequences were 0.980 and 0.864 respectively.

Table 5 shows the classification results for the pathway set. For both of the learned sequences and the test sequences, P-SOMs using CCs of 1 to 4 shifts performed best. For the learned sequences, almost all sequences were successfully classified, however for the test sequences, less than one-fourth of the sequences were classified, because the map was very complicated. It is considered to be

Table 4. Rates of successful classification for each species

name	Learned sequences	Test sequences
hsa	0.803	0.521
cfa	0.564	0.121
mmu	0.809	0.618
dme	0.967	0.962
eco	0.994	0.990
osa	0.910	0.876

Table 5. Rates of successful classification of pathway set

Input vector	CC-4	CC-4	F-4	F-4
Length	64	64	256	256
Method	P-SOM	SOM	P-SOM	SOM
Learned sequences	0.999	0.985	0.836	0.938
Test Sequences	0.240	0.208	0.214	0.195

difficult to classify the genes from different pathway sets of single organism using the features of frequencies of N-tuples or CC of sequences.

5 Conclusion

We proposed a preprocessing method for DNA sequences by using correlation coefficients of the occurrence of the nucleotides. Using this method, the clustering results of the sequences were nearly compatible with those obtained using the frequencies of the N-tuples despite the difference in the length of input vectors. The correlation coefficients are considered a more effective method for preprocessing DNA sequences.

Pareto learning SOM method is applied to the classification of DNA sequences by using correlation coefficients and frequencies as input vectors. Pareto learning SOM using CC as the input vector shows good performance for classification compared with that obtained with conventional SOMs, and frequencies. Correlation coefficients are effective as indexes for classifiertion.

In the future studys, we must apply this method to additional types sequence data, such as coding region and non-coding region, and to large data sets such as whole genomea. For such experiments, we must improve the computational costs of P-SOMs, which are 5 times more than those of conventional SOMs.

References

1. illuminal, An Introduction to Next-Generation Sequencing Technology, `http://www.illumina.com/Documents/products/Illumina_Sequencing_Introduction.pdf`
2. Kohonen, T.: Self Organizing Maps. Springer (2001)
3. Abe, T., Ikekura, T., et al.: Informatics for unreveiling hidden genome signatures. Genome Res. 13, 693–702
4. Dozono, H., Kabashima, T., et al.: Visualization of the Packet Flows using Self Organizing Maps. WSEAS Transactions on Information Science & Applications 7(1), 132–141 (2010)
5. Sorjamaa, A., Corona, F., Miche, Y., Merlin, P., Maillet, B., Séverin, E., Lendasse, A.: Sparse Linear Combination of SOMs for Data Imputation: Application to Financial Database. In: Príncipe, J.C., Miikkulainen, R. (eds.) WSOM 2009. LNCS, vol. 5629, pp. 290–297. Springer, Heidelberg (2009)
6. Dozono, H., Nakakuni, M.: Application of Supervised Pareto Learning Self Organizing Maps to Multi-modal Biometric Authentication. IPSJ Journal 49(9), 3028–3037 (2008) (in Japanese)