Mapping of the Genome Sequence Using Two-stage Self Organizing Maps

Hiroshi Dozono and Takeshi Takahashi
Faculty of Science and Engineering Saga University
1-Honjyo Saga 840-8502 JAPAN
email: ¥{hiro, takahasi}@dna.ec.saga-u.ac.jp

Keywords: Bio-informatics, Genome sequence analysis,

Abstract— In this paper, we introduce an algorithm of Self-Organizing Maps(SOM) which can map the genome sequence continuously on the map. The DNA sequences are considered to have the special features depending on the regions where the sequences are taken from or the gene functions of the proteins which are translated from the sequences. If the hidden features of the DNA sequences are extracted from the DNA sequences, they can be used for predicting the regions or the functions of the sequences. In this paper, we propose the algorithms using two stage SOM which organizes the sequences of the specific length at the first stage and organizes the set of sequences at the 2nd stage This algorithm can map the genome sequences on the map at each stage depending on the features of the sequences. In this paper, we improve the performance of the algorithm of two stage SOM using the batch updating method based on simulated annealing method to organize the adjacent sequences closely on the map at first stage.

We made some analyses of the genome sequences concerning the functions, species and secondary structure of the sequences.

.

1 Introduction

Recently, the advancement of bioinformatics is remarkable after the analysis of human genome is completed. But, not all of the functions of the protein coded from the genome are not understand. It is necessary to compare the sequence of amino-acid of unknown proteins with the sequence of the protein to have already known the function for the prediction of the function. BLAST is the algorithm which compares a given sequence with the database of proteins and it can find the similar sequence above a certain score, but the global relations among the sequences can not be extracted.

Self Organization map (SOM)[1] designed by Konohen is used for clustering, classification, and visualization of multi-dimensional data. SOM has already been used for the visualizations of genome data of DNA sequences and it can visualize the global relation of genome data, relation among the species and relation among the functions[2][3][4][5]. In this paper, we propose an algorithm using two stage SOM which maps the genome sequences on two layered maps. Two stage SOM is a kind of multi-layered SOM. As the multilayer SOM, hierarchical SOM, which processes the data layer by layer to precise the abstraction level, is widely used for image processing, signal processing and document clustering[6][7][8]. In this paper, we used the simply layered SOM which resembles multilayer perceptron. This type of 2 stage SOM uses 2nd layer for the improvement of clustering performance or mapping of the time series data extracted on the 1st layer[9][10][11](The references[9][10] are available on IEEE Xplore : URL:http://ieeexplore.org). The first layer maps the data at each sampling time. The second layer maps the series of coordinates of mapped data on the first layer. Using this algorithm, the time series is mapped as the trace of the mapped data on the second layer. In this paper, we improved the algorithm for sequence analysis.

For the first layer, we used the algorithm based on the algorithm reported at WSOM05[4]. This algorithm was aimed for mapping the DNA sequences continuously on the map using the batch update method based on simulated annealing. In this paper, we improved this algorithm as to map the sequence more efficiently and as to map the DNA sequences and the amino-acid sequences. The second layer is used to map the DNA sequence more continuously. Additionally, 2nd layer compresses the data mapped on the 1st layer. Thus, the difference among long sequences and difference of groups of the sequences can be visualized more clearly. For the second layer, we made slight changes from the algorithm reported in [11].

In this paper, we made some experiments of mapping the genome sequences classified by metabolic pathways and species, and the amino-acid sequences labeled by secondary structures of each amino acid.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

2 Analyses of the genome sequence by SOM

2.1 Genome Sequences

Recently, many genome sequences are determined and opened to the public on the WEB. The genome sequences are categorized to DNA sequences and amino-acid sequences. DNA sequence is the sequence of 4 kinds of deoxyribo nucleic acids, which are represented in 4 letters, “A”, “C”, “G” and “T”. All DNA sequence of human was already determined and about 30000 genes, which were the functional sub-sequences on DNA sequence, were found, but the functions of all genes are not determined yet. The genes are transcribed to mRNA and pre-processed and translated to amino-acid sequences, which are the sequence of 20 kinds of amino-acid represented in one of the alphabet letter. The amino-acid sequences, which work as proteins or enzymes, are also post-processed and form the 3D-structures. The 3D structures determine the functions of the sequence. The other part of the DNA sequences are also transcribed and they work as functional RNA (eg, t-RNA, r-RNA) , or siRNA and miRNA which control the expressions of genes.

As mentioned, the genome sequences are mere the sequence of strings. But, the sequences itself has meaning. For example, the DNA sequences are classified to the functional regions of genes, promoter regions of genes and junk region. The genes are classified to the metabolic pathways, e.g. amino acid pathway, energy metabolism and carbohydrate metabolism. Each gene has the function concerning the 3D structure of the protein translated from the mRNA transcribed from the DNA sequence. Furthermore, the sequences leave the trace of evolutions, so it will be possible to extract the evolutional relation among the species from the genome sequences. In the conventional sequence analysis, the sequences are mainly analyzed by using 1-dimensional information. For examples, motifs of known features are used for finding the specific regions of sequences, alignments of the known sequences and target sequences are carried out to identify functions of target sequence and hidden Markov model is used as the stochastic model of 1-dimensional sequence. The relations among DNA sequences will be more understandable if they are mapped appropriately on 2-dimentional plane.

For this purpose, we used SOM because SOM can organize the generic feature of DNA sequences by sufficient learning of known DNA sequences on 2-dimensional plane. We developed an algorithm which trains self organizing map that organizes the DNA probes on the map. The DNA probes are short DNA sequences which are strings of fixed length comprised of discrete values 'A','C','G','T'. The details of this algorithm are reported in [2]. And we improved this algorithm as to map the adjacent sequences closely on the map. For this purpose, we changed the method for searching winner units as to search the winner from adjacent units and as to update the neighboring units taking accounts the shift between the sequences arranged to the winner unit and neighboring units[3]. Furthermore, we propose the batch update method based on simulated annealing to improve the ability of arranging adjacent sequences to adjacent units[4]. Using this method, more than half of the sequences are mapped adjacently on the map. But, concerning the visualization of the sequence, a sequence is mapped as many fragments of short sequences on the map and concerning the clustering of the sequences classified by metabolic pathways and by species, each group did not mapped as rather the fragments of the regions than a region.

In this paper, we made further improvements for this algorithm and use this algorithm for the 1st stage of 2 stage SOM. The details of the improvements are mentioned in next section.

3 Mapping of the genome sequences by 2 stage SOM

3.1 2 stage SOM

2 stage SOM, which is referenced in this paper, is a kind of 2 layered SOM and used for the improvement of clustering performance and time series analysis[9][10][11] . This type of 2 stage SOM is reported at JAPAN SOM meeting held in 2007[11]. In this report, 2 stage SOM is used for the analysis of the time series of voice. The summary of this algorithm is as follows. The 1st layer is used for the learning of Fourier power spectrum of each time. After the learning of 1st layer, 2nd layer is organized. 2nd layer is used for the learning of the series of the coordinates of the mapped data on the 1st layer. The coordinate is defined by using the position of the units on two dimensional map and the series of the coordinates in pre-defined time interval are used for the learning. After learning, the time series are mapped continuously on the map.

3.2 Application of 2 stage SOM for the learning of genome sequence

We modified the 2 stage SOM for the learning of genome sequence data. For the 1st layer, we used the modified version of the sequence mapping algorithm mentioned in the previous section. At first, the input sequences are extended from DNA sequences to both of the DNA sequences and amino acid sequence. Secondly, the vector values assigned to each unit are changed from the discrete

alphabets to the probability of each alphabet at each position. This modification allows more precious expression of the sequences on each unit. Thirdly, the simulated annealing method is changed to the simulated annealing method with mean field approximation. The modification of the probabilistic expression made it easier to apply mean field approximation and the speed of the learning is highly improved. The second layer learns the series of the coordinates. After the learning of genome sequences on the 1st layer, the learned sequences are mapped and the series of coordinates of pre-defined intervals are learned on the 2nd layer. The algorithm is slightly modified as to map the consecutive sequences to the adjacent units of previous winners. The algorithm is as follows.

Algorithm of 1st layer

Step1: Initialization of the map and parameters
Initialize all of the probability P^{ij}_{kl} randomly where P^{ij}_{kl} is the probability of l-th alphabet at k-th position of the sequence of length L assigned to the unit U1(i,j). Set neighboring region range NR1=INR1, DNR1=INR1/LOOP1, threshold of global matching TH1=ITH1, DTH1=(ITH1-TTH1)/(LOOP1+LOOP2), threshold of local matching TH2=ITH2 and DTH2=(ITH2-TTH2)/(LOOP1+LOOP2).
Repeat the following steps in (LOOP1+LOOP2) steps with decreasing NR, TH1 and TH2 by DNR, DTH1 and DTH2 respectively.

Step2: Batch Learning Phase
Map all reference sequences on the map and set the counter.

Step2.1: Initialize the counter $C^{ij}_{kl} = 0$, where C^{ij}_{kl} is the number of occurrence of l-th alphabet at k-th position of the sequence assigned to the unit U1(i,j).

Step 2.2: Mapping of a genome sequence
For each reference sequence rseq repeat the following steps.

Step 2.2.1: Set the position on the reference sequence P=0.

Step 2.2.2: From all of the units, search for the winner unit W_U1(i,j) which is closest to the sub-sequence of length L start from P (input vector of 1st stage SOM) on reference sequence. The distance between U1(i,j) and sub-sequence is defined as follows.

$$dist = L - \sum_{k=1}^{L} P^{ij}_{kl}, l = rseq[P+k]$$

Step 2.2.3: If the distance between W_U1(i,j) and sub-sequence is greater than TH1, Set P=P+1 and go to step 2.2.2.

Step 2.2.4: Update the counter C^{ij}_{kl} as follows.
For each k, increment C^{ij}_{kl} where l=rseq[P+k].

Step 2.2.5: If P>(Length of rseq)-L then proceed to next reference sequence. If all reference sequences are processed , goto Step 3.

Step 2.2.6: Set P=P+1. From the adjacent units of previous winner W_U(i,j) , search for the new winner unit W_U1(i,j) which is closest to the sub-sequence of length L start from P on reference sequence.

Step 2.2.7: If the distance between W_U1(i,j) and sub-sequence is greater than TH2, go to step 2.2.2 (repeat the global search) else goto Step 2.2.4

Step 3: Batch update phase
Update P^{ij}_{kl} according to the counter value C^{ij}_{kl} .

Step 3.1: Set the temperature Temp=ITEMP, DT=log (TTEMP/ITEMP)/LOOP_N where ITEMP, TTEMP and LOOP_N is the initial temperature, terminate temperature and number of iteration in simulated annealing.
Repeat the following steps with decreasing temperature by TEMP=TEMP/exp(DT).

Step 3.2: Calculate the counter value by adding the counter values of neighboring units.
For each unit U1(i,j) execute the following steps.

Step 3.2.1: Set $C'^{ij}_{kl} = C^{ij}_{kl}$ for each k,l.

Step 3.2.2: For all units U1(n,m) in the range of neighboring U1(i,j) execute the following steps.

Step 3.2.2.1: Calculate the optimal number of shift SN by maximizing the following distance.

$$dist = \frac{\sum_k \sum_l P^{ij}_{kl} \bullet P^{nm}_{(k+SN)l}}{L - SN} \text{ where } -\frac{L}{2} \le SN \le \frac{L}{2}$$

Step 3.2.2.2: Update the counter C'^{ij}_{kl} as follows.
For each k,l, update $C'^{ij}_{kl} = C'^{ij}_{kl} + fn(d) * P^{nm}_{(k+SN)l}$ where fn(d) is the neighbor function and d is the distance between U1(i,j) and U1(n,m) on the map.

$$d = \sqrt{(n-i)^2 + (m-j)^2}$$

Step 3.3: Update P^{ij}_{kl} using simulated annealing method with mean field approximation.

$$P^{ij}_{kl} = \frac{\exp(-KC \bullet (Cm - C^{ij}_{kl}) / Temp)}{\sum_l \exp(-KC \bullet (Cm - C^{ij}_{kl}) / Temp)}$$

where $Cm = \max_l C'^{ij}_{kl}$

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

In Step 2.2.2, the distance between the units and sub-sequences are defined using rather the probability values P^{ij}_{kl} than the deterministic discrete alphabet assigned to the unit. In Step 2.2.3, the distance between the global winner unit and the reference sequence is limited in the threshold TH1 not to map the inadequate sequence to the unit. In Step 2.2.7, the distance between the adjacent winner unit and the reference sequence is also limited to avoid the excessive mapping to adjacent units. If the applicable winner can not be found from the adjacent units, it return to Step 2.2.3 to search for the global winner. In Step 3.2.2.1, the counter value C^{ij}_{kl} of the neighboring units are integrated taking accounts the shift of the sequences assigned to itself and its neighbors.

For the mapping of the genome sequence after learning, the procedure defined in Step 2 is used with setting TH1=TTH1 and TH2=TTH2.

After the learning of 1st layer, following 2nd layer algorithm is used to organize 2nd layer.

2nd layer Algorithm
Step 1: Initialization of the map and parameters.
Initialize the x^{ij}_k and y^{ij}_k (0<k<=L2) which is the vector assigned to the U2(i,j) on the 2nd layer using random value from 0 to SIZE1, where SIZE1 is the map size of layer 1. Set neighboring region range NR2=INR2, DNR2=INR2/LOOP1, learning rate LR=ILR, DLR=ILR(LOOP1+LOOP2), threshold of local matching TH3=ITH3, DTH3=(ITH3-TTH3)/(LOOP1+LOOP2) ,
Repeat the following steps in (LOOP1+LOOP2) steps with decreasing NR, LR and TH3 by DNR, DLR and DTH3 respectively.

Step 2: Batch learning phase

Step 2.1: Initialization of the update values
Initialize all of the $u\_x^{ij}_k$ and $u\_y^{ij}_k$ to 0.

Step 2.2: For each reference sequence repeat the following steps.

Step 2.2.1: Set the position on the reference sequence P=0, search mode SM=”G” and sequential number N=1.

Step 2.2.2: Search for the winner unit W_U2(i,j) which is closest to the sub-sequence starting from position P using the procedure defined in Step 2 of layer 1 algorithm.

Step 2.2.3: set Xp[N]=i, Yp[N]=j and N=N+1
If N<L2 goto Step 2.2.2
Xp and **Y**p become the vectors comprised of the coordinates of the units on the 1st layer and are used for the input vectors on the 2nd layer.

Step 2.2.4 If SM=”G” execute Step 2.2.4.1 else execute Step 2.2.4.2

Step 2.2.4.1: Search for the winner unit W_U2(k,l) which is assigned to the closest vector to **Xp** and **Yp** from all of the units on layer 2. Set SM=”L”

Step 2.2.4.2 Search for the winner units W_U2(k,l) which is assigned to the closest vector to **Xp** and **Yp** from adjacent units of previous winner**.** If sum of the distance between (Xp[k],Yp[k]) and (x^{ij}_k, y^{ij}_k) is greater than TH3, where $\mathbf{x}^{ij}=(x^{ij}_1, \ldots, x^{ij}_{L2})$ and $\mathbf{y}^{ij}=(y^{ij}_1, \ldots, y^{ij}_{L2})$ are the vectors assigned to U2(i, j). set SM=”G” and goto Step 2.2.2

Step 2.2.5: For the W_U2(k,l) and its neighboring units ranged in NR, set $u\_x^{nm}_h=u\_x^{nm}_h+fn2(d)*Xp[h]$, $u\_y^{nm}_h=u\_y^{nm}_h+fn2(d)*Yp[h]$ (0<h<=L2) where fn(d) is the neighbor function. Shift **Xp** and **Yp** as follows.
Xp[h-1]=Xp[h], Yp[h-1]=Yp[h] (1<h<=L2) and set N=N-1. Set P=P+1 and goto Step 2.2.2 until P=LENGTH(rseq)-L

Step 3: Batch update phase
For each i,j,k, update $x^{ij}_k=(1-LR)*x^{ij}_k+LR*u\_x^{ij}_k$, $y^{ij}_k=(1-LR)*y^{ij}_k+LR*u\_y^{ij}_k$

In step 2.2.4, the distance measure of the sub-sequence based on the coordinates on the 1st layer is used. This comes from the assumption that similar sequences are mapped closely on the 1st layer map. In this paper, we used the algorithm based on simulated annealing in 1st layer and it improves the availability of this assumption compared with simple SOM algorithm. In step 2.2.4.2, the threshold value TH3 is used to avoid the excessive mapping to adjacent units. The length of the vector L2 on the 2nd layer is set longer than L1 which is the length of vector on the 1st layer. Thus, features of longer sequences are extracted on the 2nd layer. Compared with hierarchical SOM, the abstraction level between the layer becomes inversely, that is, more abstracted features of longer sequences are extracted in the 2nd layer.

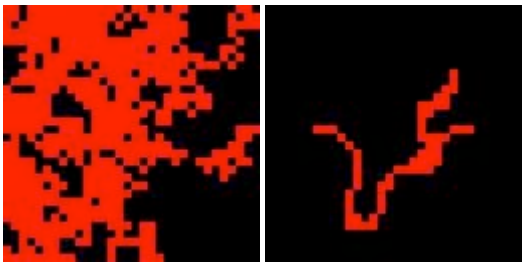
4 Experimental results

4.1 Experimental settings

We made some experiments of mapping genome sequences classified by metabolic pathways and species, and the amino-acid sequences labeled by secondary structures of each amino acid. The sequence data of metabolic pathways are taken from KEGG database[12] and the sequence data labeled by secondary structures are taken from DSSP database[13].

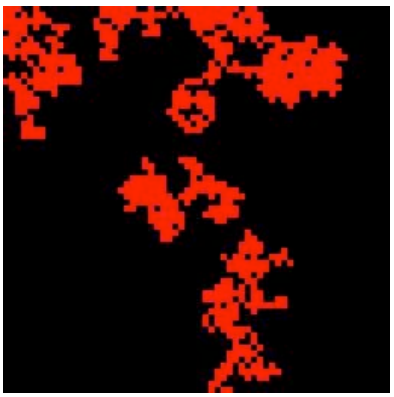
4.2 Mapping of DNA sequences from metabolic pathway

At first, the experimental results for mapping the DNA sequences from metabolic pathways. 1008 DNA sequences coding the metabolic pathways of, amino acid metabolic pathways, chromosome metabolic pathways, energy metabolic pathways, nucleotide metabolic pathways, metabolism of complex lipids and Signal transduction are used for learning. Figure 1 shows the mapping of a sequence from amino acid metabolism numbered 162417 in KEGG database.

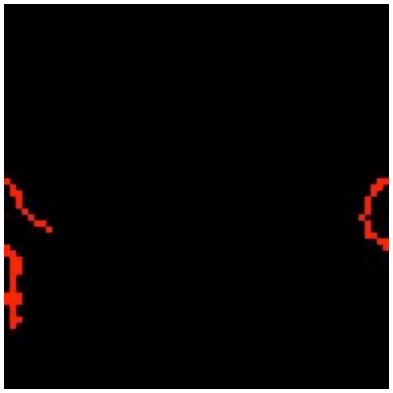


Layer 1 map Layer 2 map
Figure.1 Mapping of the DNA sequence 162417

The parameters of SOM are set as map size of layer 1 and layer 2 is 32, L=6, L2=12, INR1=2, INR2=4, LR=0.8 TH1=5, TH2=4, TH3=4, LOOP1=50, LOOP2=5, LOOP_N=10, KC=1.0, TI=0.1, TT=0.001 and the length of the sequence is 1632. Over the 99% of the sub-sequence and set of the subsequences are mapped on layer 1 and layer 2 respectively under this condition and over 99% of the sequences are mapped adjacently on both maps. On layer 1, too many units are matched to the sequences, but in layer 2 the sequence are mapped on the region comprised of some curved lines. Layer 2 is considered to be compressing the mapping of layer 1. Other sequences are also mapped in the same manner.



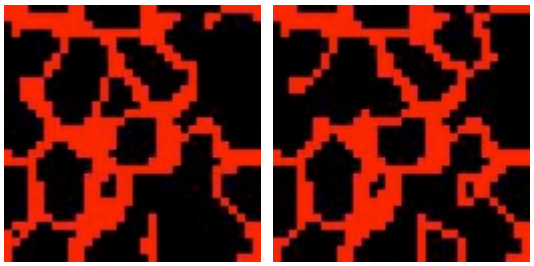
Layer 1 map



Layer 2 map
Figure 2 Mapping of DNA sequence 162417 on the map sized 64x64

Next, the mapping results with changing the map size to 64x64 are shown in figure 2. On layer 1, the sequence is mapped to some connective regions. In these experiments, the circular maps are used. Considering that, on layer2, the sequence is mapped as a curved line.

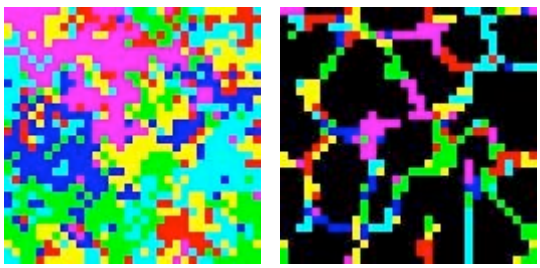
Next, the mapping results on layer 2 of all genes concerning amino acid metabolic pathways and energy metabolic pathways on the map sized 32x32 are shown in figure 3.



Amino acid metabolism Energy metabolism
Figure 3 Mapping of metabolic pathways

On layer 1, all of the units have the mapped sub-sequences. The Mapping results are almost same, but some parts are different. The sequences mapped on the different regions denote the differences between these metabolic pathways.

Figure 4 shows the mapping result of all learned metabolic pathway on the map sized 32x 32.

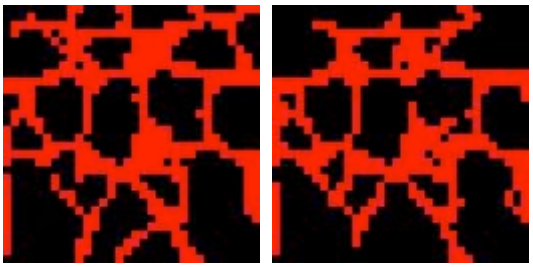


layer 1 map layer 2 map
Figure 4 Mapping of all learned metabolic pathways

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

Each color denotes the each metabolic pathway. The layer 1 map shows good separations among the metabolic pathways. Compared with the algorithm reported in [4], the mapping results are much improved. But, Layer 2 map shows the many fragmented region concerning each metabolic pathway. For this purpose, layer 1 map shows adequate results and layer 2 map is considered to be not applicable.

Next, we made other experiment concerning the genes taken from different species. 135 genes taken from amino acid metabolic pathways of 5 species which include the words "amino" and "acid" in the keywords of PATHWAY are selected from KEGG database and used for learning. Figure 5 shows the mapping results of all learned genes of Homo Sapience and Drosophila Melangaster on layer 2.

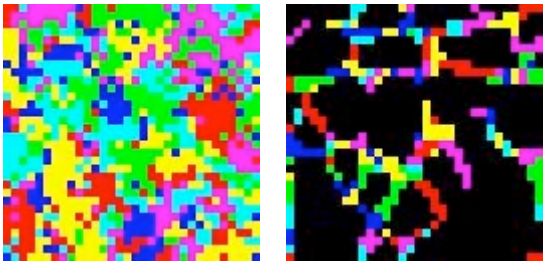


Map of Homo Sapience Map of Drosophila Melangaster

Figure 5 Mapping of the genes of different species

Alike a results shown in Figure 3, the difference of the usage of 6-tuples of DNA sequences of the genes is shown as the difference of the mapping results.

Figure 6 shows the results of mapping of the genes from all learned species.

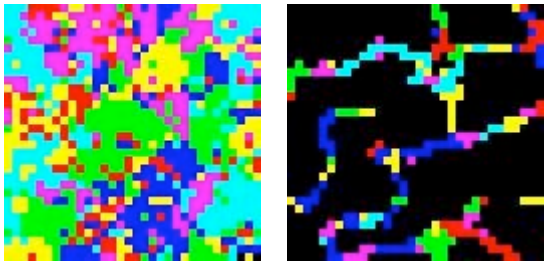


Layer 1 map Layer 2 map

Figure 6 Mapping of the genes from 5 species

Alike figure 3, Layer 1 map shows rather separated regions, but the layer 2 map is too much compressed.

Next, we made experiments using amino acid sequence data. At first, the amino acid sequences of amino acid metabolic pathways of 5 species used in the previous experiments are used for learning and mapped on the organized map.

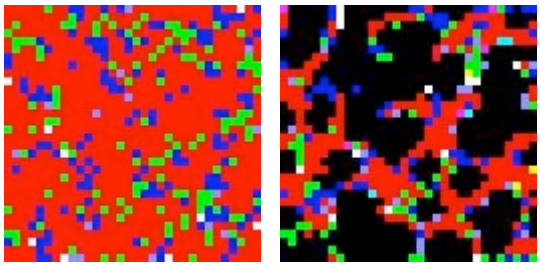


Layer 1 map Layer 2 map

Figure 7 Mapping of amino acid sequences of the genes from 5 species

For amino acid sequence, the parameters are set as L=4, TH1=3, TH2=2, because the variation of the alphabets are increased from 4 to 20. Layer 1 map shows better separation compared with the map shown in figure 6. This result shows that this algorithm is also applicable to amino acid sequences.

At last, we made the experiments of the mapping the amino acid sequence data labeled by secondary structures. Figure 8 shows the results.



Layer 1 Map Layer 2 Map

Figure 8 Mapping of the secondary structures

In this figure, red, blue, green points denote the secondary structure helix, beta-sheet and Coil respectively. According to the protein secondary structure, helix has large majority, thus almost of the map are colored in red. On layer 2 map, the region of the helix is compressed, so the minor structures are be enhanced relatively.

5 Conclusions

We developed the sequence mapping algorithm using 2 stage Self Organizing Map and made some experiments of mapping the DNA sequences and amino acid sequences classified by metabolic pathways and species. We made further improvements to the 1st layer algorithm using simulated annealing in batch update phase. Using this algorithm, the sequences are mapped in the continuous region on both layers and the 2nd layer shows the compressed map which corresponds to the feature of the sequence or the set of sequences. But, 2nd layer map is

too much compressed to examine the relation between the set of sequences.

As the future work, we must examine the biological meaning of the maps and must improve the algorithm not to organize the excessively compressed map on 2nd layer.

References

[1] T. Kohonen: Self Organizing Maps, Springer, ISBN 3-540-67921-9

[2] Hiroshi Dozono, A Design Method of DNA chips for SNP Analysis Using Self Organizing Maps, Advances in Self-Organizing Maps, Springer, 4, 152-159, (2001)

[3] Hiroshi Dozono, A Design Method of DNA chips Using Hierarchical Self Organizing Maps, Proceedings of WSOM'03,(2003)

[4] Hiroshi Dozono, An Algorithm of SOM using Simulated Annealing in the Batch Update Phase for Sequence Analysis, Proceedings of 5th Workshop on Self Organizing Maps, pp. 171--178 (005)

[5] Abe,T., Ikemura, T., Kanaya, S.: Kinouch, S. and Sugawara, H.,A Novel Bioinformatics Strategy for Phylogenetic Study of Genomic Sequence Fragments: Self Organizing Map(SOM) of Oligonucleitide Frequencies, Proceedings of 5th Workshop on Self Organizing Maps, pp. 669--676 (005)

[6] Koh,J. Suk, M. and Bhandarkar S., "A Multilayer Self-Organizing Feature Map for Range Image Segmentation", Neural Networks, Volume 8, No.1, pp.67-86, 1985

[7] Shalash, W.M.; Abou-Chadi, F. , A fingerprint classification technique using multilayer SOM, Radio Science Conference, 2000. 17th NRSC apos;2000. Seventeenth National, Volume,Issue, 2000 Page(s):C26/1 - C26/8

[8] J. Lampinen and E. Oja. Clustering properties of hierarchial self-organizing maps. Journal of Mathematical Imaging and Vision, 1992.

[9] Koike, K. Kato, S. Horiuchi, T. A two-stage self-organizing map with threshold operation for data classification, SICE 2002. Proceedings of the 41st SICE Annual Conference, Volume: 5, On page(s): 3097- 3099 vol.5

[10] Satoru Kato , Kenta Koike , Tadashi Horiuchi , A study on two-stage self-organizing map and its application to clustering problems, Electrical Engineering in Japan, Volume 159, Issue 1 , Pages 46 – 53

[11] Kazuhiro Kotetsu,et.al: Suggestion of Specification Signal Identification Method for Time Series Signal Using Self Organizing Map, Technical Reports of 8th Annual Meeting of Self Organizing Maps in JAPAN 2007, pp.39-44

[12] Kyoto Encyclopedia of Genes and Genomes: http://www.genome.ad.jp/kegg

[13] The DSSP software and database : http://swift.cmbi.ru.nl/gv/dssp/