DNA Research **16**, 287–298, (2009) doi:10.1093/dnares/dsp018

# A Novel Bioinformatics Strategy for Function Prediction of Poorly-Characterized Protein Genes Obtained from Metagenome Analyses

Takashi Abe[1,*], Shigehiko Kanaya[2], Hiroshi Uehara[1], and Toshimichi Ikemura[1]

*Nagahama Institute of Bio-Science and Technology, Tamura-cho 1266, Nagahama-shi, Shiga-ken 526-0829, Japan[1] and Department of Bioinformatics and Genomes, Graduate School of Information Science, Nara Institute of Science and Technology, Takayama, Ikoma, Nara 630-0101, Japan[2]*

(Received 16 July 2009; accepted 3 September 2009; published online 3 October 2009)

**Abstract**

**As a result of remarkable progresses of DNA sequencing technology, vast quantities of genomic sequences have been decoded. Homology search for amino acid sequences, such as BLAST, has become a basic tool for assigning functions of genes/proteins when genomic sequences are decoded. Although the homology search has clearly been a powerful and irreplaceable method, the functions of only 50% or fewer of genes can be predicted when a novel genome is decoded. A prediction method independent of the homology search is urgently needed. By analyzing oligonucleotide compositions in genomic sequences, we previously developed a modified Self-Organizing Map 'BLSOM' that clustered genomic fragments according to phylotype with no advance knowledge of phylotype. Using BLSOM for di-, tri- and tetrapeptide compositions, we developed a system to enable separation (self-organization) of proteins by function. Analyzing oligopeptide frequencies in proteins previously classified into COGs (clusters of orthologous groups of proteins), BLSOMs could faithfully reproduce the COG classifications. This indicated that proteins, whose functions are unknown because of lack of significant sequence similarity with function-known proteins, can be related to function-known proteins based on similarity in oligopeptide composition. BLSOM was applied to predict functions of vast quantities of proteins derived from mixed genomes in environmental samples.**

**Keywords:** batch learning SOM; oligopeptide frequency; protein function; metagenome; alignment-free clustering

## 1. Introduction

A large number of protein genes whose functions cannot be predicted by the sequence homology search have progressively accumulated in the International DNA Sequence Databanks and remain of very little use to science and industry. Currently, the situation has become worse, since the number of proteins with unknown functions, such as those found in novel unculturable species derived from environmental samples, has increased rapidly. Considering that such annotation is done based on protein sequence alignment, other methods that are complementary to the sequence alignment would be required. One approach for the prediction of function is detection of similarity in three-dimensional (3D) structure of proteins, by direct comparison of coordinates or threading approach. Another successful approach is an integration of 3D structure information with one-dimensional sequence similarity, and the programs and the structural domain databases (e.g. FUGUE, SCOP and CATH) have been published.[1–5] Protein sequence alignments in twilight-zone of similarity have also been studied intensively.[6–8] As described earlier, however, the number of proteins with unknown functions, such as those

Edited by Katsumi Isono

* To whom correspondence should be addressed. Tel. +81 749-64-8126. Fax. +81 749-64-8126. E-mail: takaabe@nagahama-i-bio.ac.jp

found in novel unculturable species derived from environmental samples, has increased rapidly, and in light of technical limitations and the cost required, it is difficult to apply routinely the higher-order structure determination of proteins for prediction of functions of the increasingly vast quantity of such novel proteins.

In the present study, we have developed a sequence alignment- and structural information-free clustering method that will complement the function prediction for proteins for which no or only ambiguous prediction could be obtained even by a combination of conventional tools. Self-organizing map (SOM) is an unsupervised clustering algorithm developed by Kohonen *et al*.,[9,10] which enables us to efficiently and easily interpret the clustering of high-dimensional data utilizing visualization on a two-dimensional plane. More than 15 years ago, Ferran *et al*.[11] performed pioneering and extensive SOM analyses for dipeptide compositions in ~2000 human proteins stored in the SwissProt Database. They reported clustering (self-organization) of the proteins by both function and higher-order structure with no advance knowledge during the SOM calculation. Although this unsupervised clustering method can be considered useful for predicting protein function, the study was conducted long before decoding of genome sequences, and at that time proteins of unknown function were rare. Furthermore, because a long computation time was required for the conventional SOM of the dipeptide compositions (400-dimensional vectorial data), even using high-performance computers at that time, and because the final map was dependent on both the order of the vectorial data input and the initial vectorial values, the SOM method has rarely been used for prediction of protein function.

Previously, we developed a modified SOM (batch-learning SOM: BLSOM), which depends on neither the order of data input nor the initial condition, for codon frequencies in gene sequences[12] and oligonucleotide frequencies in genomic sequences.[13] The BLSOM recognizes species-specific characteristics of codon or oligonucleotide composition, permitting clustering (self-organization) of genes and genomic sequences according to species without the need for species information during BLSOM calculation. Various high-performance supercomputers are now available for biological studies, and the BLSOM, although not the conventional SOM, is suitable for actualizing high-performance parallel-computing[14–16] and thus for a large-scale computation of a huge quantity of data using a high-performance supercomputer. In our previous studies, we used the BLSOM for phylogenetic classification of genomic sequence fragments obtained from mixed genomes of environmental microorganisms by analyzing tetranucleotide frequencies (256-dimensional vectorial data).[17–20] The unsupervised method clustered the genomic fragment sequences (e.g. 10 kb) according to phylotype without phylotype information.

For the present study, we developed BLSOM to predict protein function on the basis of similarity in oligopeptide composition (di-, tri- and tetrapeptide compositions in this study) of proteins. We searched the BLSOM condition to faithfully reproduce separation (self-organization) of function-known proteins according to functional categories. As a test data set for this purpose, we focused on protein genes which were classified into COGs (clusters of orthologous groups of proteins) by an NCBI group.[21] The NCBI COG (abbreviated as NcCOG) is the functional category identified with best-hit relationships between the completely sequenced genomes using the sequence homology search. The NcCOG is not a simple, decisive categorization because of the co-orthologous relationship of protein genes produced by duplication of orthologous genes subsequent to a speciation event.[22–24] The NcCOG, however, is undoubtedly a useful categorization of proteins according to function, especially from the view point of the method-oriented purpose of the present study because proteins belonging to a single NcCOG most likely have the same function. In the first approach, by analyzing di-, tri- and tetrapeptide compositions in the 83 962 proteins classified by the NCBI group[21] into 2853 function-known NcCOGs, we could obtain BLSOM conditions that faithfully reproduced the NcCOG classifications without an advanced knowledge of the NcCOG information. Then, we applied the BLSOMs to predict functions of proteins obtained from environmental mixed genomes.

## 2. Materials and methods

### 2.1. Amino acid sequences

Amino acid sequences were obtained from http://www.ncbi.nlm.nih.gov/GenBank/. Proteins shorter than 200 amino acids in length were not included in the present study. To reduce the computation time as described by Ferran *et al*.,[11] BLSOM was constructed with tripeptide frequencies of the degenerate 11 groups of residues: {V, L, I}, {T, S}, {N, Q}, {E, D}, {K, R, H}, {Y, F, W}, {M}, {P}, {C}, {A} and {G}. BLSOM was also constructed with tetrapeptide frequencies of degenerate six groups of residues: {V, L, I, M}, {T, S, P, G, A}, {E, D, N, Q}, {K, R, H}, {Y, F, W} and {C}.[11]

### 2.2. Blsom construction

We previously modified the conventional Kohonen's SOM[9,10] for genome informatics to make the learning process and resulting map independent of the order of data input.[12,13] In the case of the conventional

SOM, the initial vectorial data are set at a random value, but this results in a final map which is changed by each initial data set. In contrast, in our BLSOM for genome analyses, we could obtain a reproducible map using the first and second primary components in the principal component analysis of oligonucleotide composition for the initial vectorial data. The initial weight vectors ($w_{ij}$) were arranged in the two-dimensional lattice denoted by $i$ (=0, 1, ..., $I-1$) and $j$ (=0, 1, ..., $J-1$). $\sigma_1$ and $\sigma_2$ were the standard deviations of the first and second principal components, respectively, and $J$ was defined by the nearest integer greater than $(\sigma_2/\sigma_1) \times I$. Under this condition, $I$ was set to attain the map size which provides a mean ca. 8 data points per lattice point, as described previously.[12,13]

When this strategy was applied to dipeptide composition in proteins in a preliminary study, the first component tended to reflect the length of proteins. Although the length of a protein undoubtedly relates to its function, the length sometimes differs significantly even between proteins with the same function. Furthermore, to predict functions of proteins with high novelty, it would become important to compare the oligopeptide composition of proteins with similar functions derived from the genomes with divergent evolutionary origins, and lengths of proteins with similar functions may differ significantly from each other. Additionally, when the protein-coding regions are estimated computationally from sequences of poorly characterized genomes, including those obtained by metagenome analyses, it is often difficult to specify accurately the initiation codon position and thus the protein length. To develop a method that is less dependent on protein length, we provided a window of 200 amino acids that is moved with a 50-amino acid step. For the final, residual sequence shorter than 200 amino acids, a 200-amino acid window was newly set from the last amino acid of the protein. The BLSOM was constructed for all these overlapped 200-amino acid sequences according to the method described for the oligonucleotide composition in genomic sequnces.[13] Introduction of the window with the sliding step enables us to analyze multifunctional and multidomain proteins, which originated often from the fusion of distinct proteins during evolution, collectively with smaller proteins. Only in Fig. 1A and Table 1, dipeptide frequencies in the full-length sequences of proteins were analyzed, resulting in a lower level of proper clustering according to the COG category.

The present BLSOM was suitable for actualizing high-performance parallel-computing with high-performance supercomputers. Using 136 CPUs of 'the Earth Simulator', calculations in this study could be performed primarily within 2 days. The BLSOM program suitable for PC cluster systems and a PC program for mapping of new sequences on a large-scale BLSOM constructed with high-performance supercomputers could be obtained from our group (takaabe@nagahama-i-bio.ac.jp) and UNTROD, Inc. (k_wada@nagahama-i-bio.ac.jp). Twenty thousands protein sequences could be mapped within a day using the PC system.

### 2.3. *Assignment of NcCOGs highly associated with each other*

To examine whether the sharing of one lattice point with different NcCOG sequences reflected similarity of function, we focused on all lattice points that contained 200-amino acid fragments from more than one NcCOG (designated as 'mixed lattice points') and then selected the lattice points that had more than four fragments at least of two NcCOGs in order

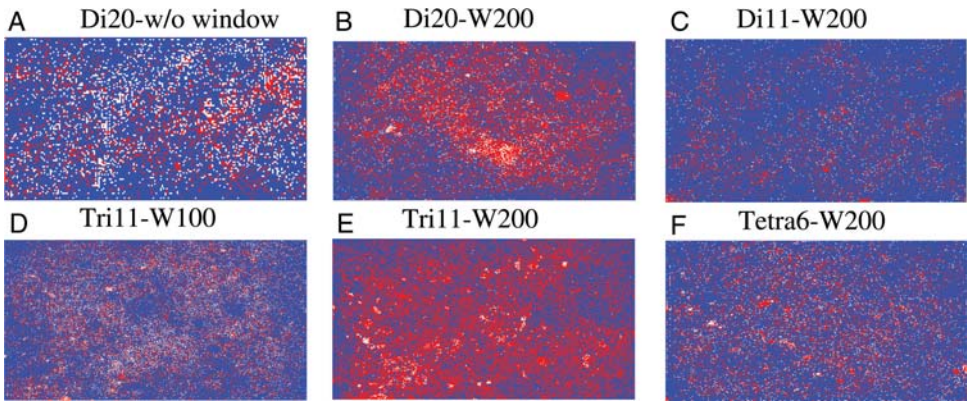


**Figure 1.** The distribution of pure lattice points. (A) Di20-w/o window; 20-amino acid groups, full-length sequences without window. (B) Di20-W200; 20-amino acid groups, 200-amino acid windows. (C) Di11-W200; 11-amino acid groups, 200-amino acid windows. (D) Tri11-W100; 11-amino acid groups, 100-amino acid windows. (E) Tri11-W200; 11-amino acid groups, 200-amino acid windows. (F) Tetra6-W200; 6-amino acid groups, 200-amino acid windows. 'Pure lattice points' are colored in red and 'mixed lattice points' are colored in blue. Lattice points without sequence or with one sequence are shown in white.

**Table 1.** Occurrence levels of pure lattice points

| Analysis condition | Proportion (%) of pure lattice points |
|---|---|
| Di20-w/o window | 15.2 |
| Di20-W200 | 35.1 |
| Di11-W200 | 10.7 |
| Tri11-W100 | 19.1 |
| Tri11-W200 | 45.4 |
| Tetra6-W200 | 18.1 |

to prevents cases of mere, accidental association; it should be noted that the number of lattice points on each BLSOM was designed to have eight data per lattice point on average. Next, for each of the mixed lattice points thus selected, we listed all pairs of NcCOGs having more than four fragments and summed up the pairs for all of the mixed lattice points. Then, for each pair of NcCOGs, the number of lattice points giving the pair was counted, and the pair found only once was omitted to reduce the case of an accidental association.

### 2.4. *Function prediction for protein-gene candidates obtained by metagenome analyses*

For each Sargasso protein, a 200-amino acid window moved with a 10-amino acid step was provided and each 200-amino acid segment was mapped on a BLSOM constructed in advance with NcCOG sequences, by identifying the lattice-point with the minimum Euclidean distances in the multidimensional space, as published previously.[14] As to every lattice point on which Sargasso fragments were mapped, the most abundant and the second abundant NcCOG segments greater than three were identified, and the mapped Sargasso segments were tentatively assumed to belong to the NcCOGs. Considering the finding that proteins belonging to different but functionally related COGs were often associated with each other on BLSOMs, the second most abundant NcCOG was included in the analysis, but the results were similar to those of the analysis in which only the most abundant NcCOG segment was chosen. Finally, when more than 60% of the number of the 200-amino acid fragments derived from one Sargasso protein gave the same NcCOG category, the Sargasso protein was assigned to that COG and designated as SomCOG, showing that the respective Sargasso protein was assigned to this COG by BLSOM. The 10-amino acid step, rather than the 50-amino acid used for the BLSOM construction, was set for this mapping process in order to reduce the position effect of the 200-amino acid segmentation. This provided five times more mapping data and reduced in part the sequence length-dependency of the prediction accuracy. Although this mapping could be conducted with PC-level computers, BLSOMs could be constructed by using high-performance supercomputers. The 10-amino acid step was not used for BLSOM construction because of the five times more computation time needed.

## 3. Results

### 3.1. *Blsoms for proteins belonging to function-known NcCOGs*

For the test data set to examine whether proteins are clustered according to function by BLSOM, we chose proteins that were classified into 2853 function-known COGs by an NCBI group[21] (NcCOGs) and analyzed oligopeptide frequencies in 83 962 proteins belonging to the function-known NcCOGs. Ferran *et al.*[11] reported, by using the conventional SOM, that the SOM for dipeptide composition after grouping of 20 amino acids into 11 categories ($11^2 = 121$ dimensional vectorial data), in which amino acids having similar physico-chemical properties were grouped as the same residue (see Materials and methods), led to a clustering that was similar to the one obtained with the 400 ($=20^2$) dimensional vectorial representation for 20 amino acids. This grouping of amino acids resulted in a substantial reduction in the computation time. To further reduce the computation time, they also attempted the amino-acid grouping into six categories ($6^2 = 36$ dimensional data; see Materials and methods) and found reduction in proper clustering into functional categories in the case of the dipeptide frequencies. Taking the pioneering findings by Ferran *et al.*[11] into account, we investigated four different BLSOM conditions to determine which gave the best accuracy and to what degree similar results were obtained among these conditions. One was the BLSOM for the dipeptide composition of 20 amino acids (abbreviated as Di20-BLSOM). The second and the third were the BLSOMs for the dipeptide and the tripeptide composition after grouping into 11 categories, 121 ($=11^2$) or 1331 ($=11^3$) dimensional data (abbreviated as Di11- or Tri11-BLSOM, respectively). The fourth was the BLSOM for the tetrapeptide composition after grouping into six categories, 1296 ($=6^4$) dimensional data (Tetra6-BLSOM). BLSOMs for much higher dimensional data such as those for the tripeptide composition of 20 amino acids (8000-dimensional data) and for the tetrapeptide composition after grouping 11 categories (14 641-dimensional data) were impossible in this study because of the limitation of resources of the supercomputers currently available to our group.

In order to introduce a method that is less dependent on the amino acid sequence length (see Materials and methods), we provided a window of 200 amino acids that is moved with a 50-amino acid step for proteins longer than 200 amino acids, and the BLSOM was constructed for the overlapped 200-amino acid sequences (a total of 472 572 sequences). Introduction of a window with the sliding step enabled us to analyze multifunctional and multidomain proteins, which originated often from the fusion of distinct proteins during evolution, collectively with smaller proteins.

The most critical part of the present study is to know the level at which each lattice point on a BLSOM contains fragments derived from a single NcCOG. The number of the function-known NcCOG categories is 2835, and the map size of a BLSOM was chosen so as to provide a mean ca. 8 data points per lattice point on the BLSOM. If sequences were randomly distributed, the probability that all fragments associated within one lattice point were derived from a single NcCOG by chance should be extremely low, e.g. $(1/2853)^8 = 2.3 \times 10^{-28}$, whereas this value should depend on the number of fragments derived from proteins belonging to the respective NcCOG. We designate here the lattice point that contained fragments derived from only one NcCOG as a 'pure lattice point'. By this definition, the lattice points that contained only one sequence fragment were not assigned to the pure lattice point, although such lattice points were minor.

Considering that the occurrence probability of the pure lattice point as an accidental event is extremely low, we next compared this occurrence level among six different BLSOMs listed in Table 1. To graphically show the difference among these BLSOMs, pure lattice points were colored as red and 'mixed lattice points', which contained fragments derived from different NcCOGs, were colored as blue (Fig. 1). Even no NcCOG category information was given during BLSOM calculation, a high percentage of pure lattice points, i.e. clustering (self organization) of proteins according to one NcCOG category, was obtained, and pure lattice points distributed all over the maps. The highest occurrence level of pure lattice points was observed on the Tri11-BLSOM for 200-amino acid fragments (Fig. 1E); ~45% of lattice points of the Tri11-BLSOM contained sequences derived from one NcCOG (Table 1). Shorting of the fragment size into 100 amino acids reduced significantly the occurrence level of pure lattice points (Fig. 1D). The dipeptide frequencies in the full-length sequences of proteins (Fig. 1A) gave a lower occurrence of pure lattice points than the dipeptide frequencies in the 200-amino acid window (Fig. 1B).

Although the results in Fig. 1 and Table 1 show that the occurrence level of pure lattice points was many orders of magnitude higher than that of random association, it was not clear in what way the 200-amino acid segments belonging to a single NcCOG were distributed on a BLSOM. In Fig. 2, the number of sequences classified into each pure lattice-point on a BLSOM was shown by the height of the vertical bar with a color representing each of the 20 NcCOG cases. These 20 examples were chosen because they were well separated from each other on a BLSOM and thus were easily visualized on one plane. Sequences belonging to a single NcCOG were localized not in a single lattice point, but in the neighboring points, resulting in a high peak composed of neighboring high bars. A few high peaks with the same color located far apart from each other were also observed. Detailed inspection showed that these detached high peaks are mostly due to the different 200-amino acid segments (e.g. anterior and posterior portions) derived from one protein, which have distinct oligonucleotide compositions and likely represented distinct structural and/or functional domains of the respective protein. This type of distinct but major peak may be informative for prediction of functions of multidomain, multifunctional proteins.

### 3.2. *NcCOGs highly associated with each other on BLSOMs*

Protein sequences contain all the following information: (i) evolutionary origin of the sequence; (ii) functional requirements (e.g. those needed to form active sites, such as functional motifs); and (iii) structural demands (e.g. the need to form a certain 3D structure or to fit into a membrane medium). Proteins belonging to a single NcCOG should have all the common information. At the same time, it should be stressed that the paralogous genes have primarily the same/similar functions. When we consider the prediction of function of the proteins derived from poorly characterized genomes such as those analyzed in metagenomic approaches, proteins that originated by diversifications of one gene during evolution and likely belong to different NcCOGs (e.g. paralogous genes) should become important because they have likely the same/similar functions. In other words, the COG category may be too strict to judge the level of clustering of proteins according to function.

Proteins with the same/similar functions such as those belonging to one paralogous group might be associated with each other on BLSOMs. To examine whether the sharing of one lattice point with different COG sequences reflected similarity of protein function, we focused on mixed lattice points, which had sequences derived from different NcCOGs, and

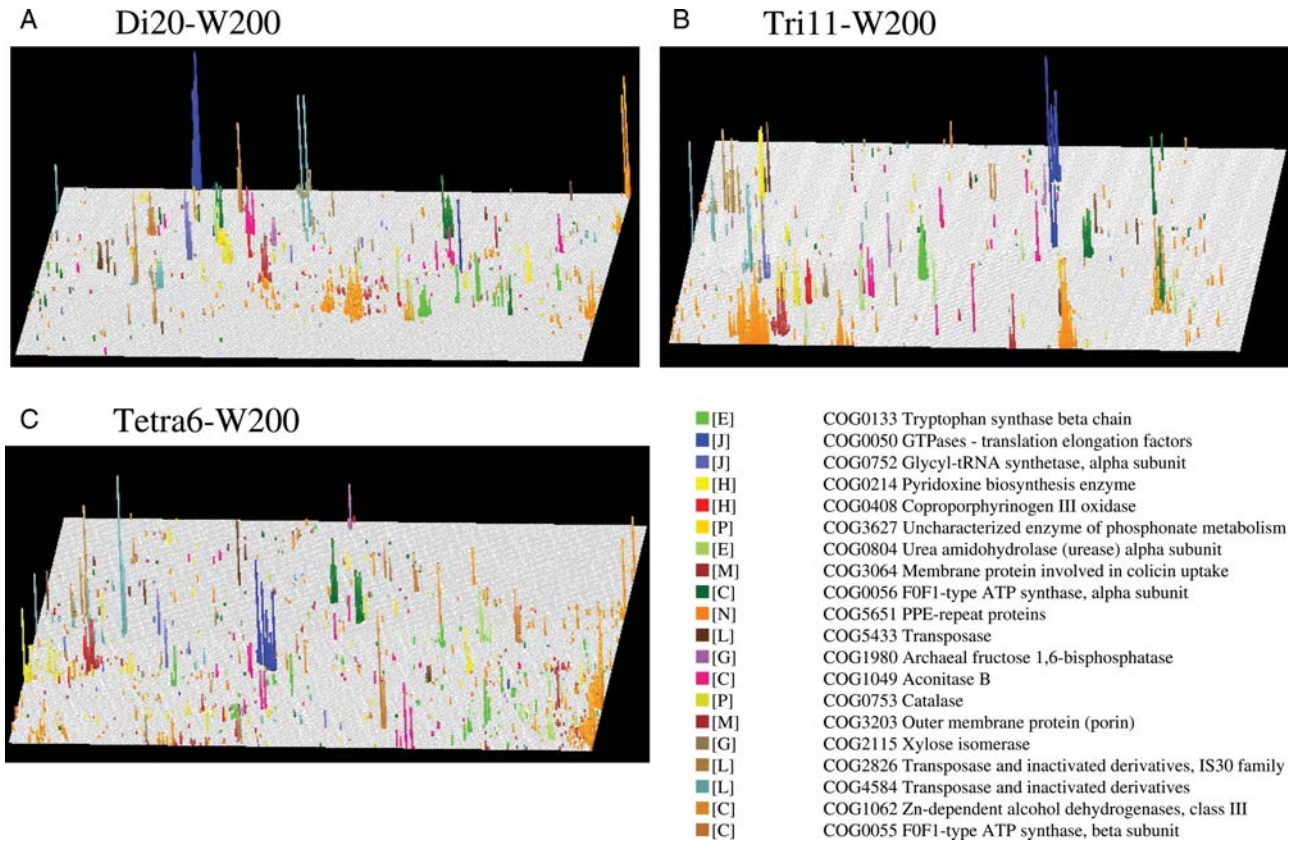


**Figure 2.** Clustering of protein sequences according to COG. (A) Di20-W200. (B) Tri11-W200. (C) Tetra6-W200. Sequences of 20 COG examples are presented. Numbers of sequences classified into each lattice point are presented by the height of the vertical bar with a color representing each of the 20 NcCOGs.

selected a set of NcCOGs whose sequences were highly associated with each on Di20-, Tri11- and Tetra6-BLSOMs, as described in Materials and methods. Table 2 lists the associated NcCOG pairs commonly found for the three BLSOMs along with the annotation of the NcCOG. When we referred to their annotations, 10 out of 11 NcCOG pairs were found to have a closely related function (e.g. ATPase, carboxylase or permease) for each pair, showing that the BLSOM can cluster proteins according to function. In the case of the exceptional one pair (COGs 0515 and 5099), a significant sequence similarity could not found. However, after fine tuning of conditions for a dot-matrix method, a weak local similarity for 10 segments was detected, whereas the segments were not linearly arranged along the primary

**Table 2.** COG pairs associated commonly on three BLSOMs

| | |
|---|---|
| COG0419 ATPase involved in DNA repair | COG0497 ATPase involved in DNA repair |
| COG0419 ATPase involved in DNA repair | COG1196 Chromosome segregation ATPases |
| COG0419 ATPase involved in DNA repair | COG4942 Membrane-bound metallopeptidase |
| COG0419 ATPase involved in DNA repair | COG5022 Myosin heavy chain |
| COG1196 Chromosome segregation ATPases | COG5022 Myosin heavy chain |
| COG0439 Biotin carboxylase | COG4770 Acetyl/propionyl-CoA carboxylase, alpha subunit |
| COG0477 Permeases of the major facilitator superfamily | COG0697 Permeases of the drug/metabolite transporter (DMT) superfamily |
| COG0477 Permeases of the major facilitator superfamily | COG2814 Arabinose efflux permease |
| COG0515 Serine/threonine protein kinase | COG5099 RNA-binding protein of the Puf family, translational repressor |
| COG1298 Flagellar biosynthesis pathway, component FlhA | COG4789 Type III secretory pathway, component EscV |
| COG3839 ABC-type sugar transport systems, ATPase components | COG3842 ABC-type spermidine/putrescine transport systems, ATPase components |

sequence. Detailed study on proteins with such local, weak similarity will be one of our future works.

### 3.3. *Mapping of function-known environmental proteins on BLSOMs constructed with NcCOG sequences*

Environmental microorganisms cannot be cultured easily under laboratory conditions, and therefore, genomes of unculturable microorganisms have remained mostly uncharacterized but are thought to contain a wide range of novel protein genes of scientific and industrial usefulness. Metagenomic approaches, which determine the sequence of mixed genomes of uncultured environmental microbes, have been recently developed to identify wide varieties of novel and industrially useful genes.[25,26] One important contribution of the present alignment-free clustering method is to systematically and efficiently predict functions of the increasingly vast quantity of function-unknown proteins derived from poorly characterized microbe genomes. To examine the feasibility of BLSOMs for function prediction for such environmental samples and compare the BLSOM results with those obtained by the sequence homology search, we next used, as a test data set, protein candidates that were obtained from genomic fragments derived from metagenome libraries originating in the Sargasso Sea[27] and examined the similarities and differences in predictions obtained by BLAST and BLSOM. We chose the Sargasso Sea sequences because the sequences covered a wide variety of genes from a wide variety of phylotypes, and in particular we focused on the 97 838 protein candidates that have been registered as ORFs in the International DNA Databanks without function annotation and are longer than 200 amino acids. From these proteins, we initially selected those that could be robustly assigned to COGs with BLAST basing on the strict criterion described by the International DNA Bank of Japan (DDBJ).[28]

It should be mentioned that, as the test data used in Fig. 1 and Table 1, we focused only on proteins belonging to function-known NcCOGs (2853 groups) to examine self-organization of proteins according to function. However, environmental sequences such as those of the Sargasso samples should have functions corresponding not only to the function-known NcCOGs, but also the function-unknown NcCOGs (2020 groups). Therefore, both types of NcCOGs (a total of 4873 groups) were included for this selection of Sargasso proteins robustly assignable to NcCOGs with BLAST.[28] A total of 4240 Sargasso proteins were assigned to the NcCOGs, and the COGs thus assigned were called HomCOGs (COGs assigned by the homology search). Then, we divided the HomCOG Sargasso proteins into 200-amino acid segments with a sliding step of 10 amino acids, obtaining a total of 77 504 segments.

As a separate analysis, we prepared the Di20-, Tri11- and Tetra6-BLSOMs with 200-amino acid segments with a sliding step of 50 amino acids, which were derived from the function-known plus-unknown NcCOGs (a total of 4873 groups: 124 292 sequences: 693 553 segments). Then, on these BLSOMs, each of the 200-amino acid segments that were derived from the HomCOG Sargasso proteins was mapped by finding the lattice point with the minimum Euclidian distance in the multidimensional space, as described in Materials and methods. For each HomCOG Sargasso protein, we summed up the mapping results of all 200-amino acid segments. If more than 60% of the mapped results gave the same NcCOG category, the Sargasso protein was assigned to this COG and this COG was designated as the SomCOG (the COG assigned with BLSOM). Because the number of COG categories used here was 4873, a probability of random chance that the multiple segments derived from one Sargasso protein were classified into the lattice points tentatively assigned to the same NcCOG should be extremely low. A Venn diagram shows high levels of overlap of the COG assignments obtained by three BLSOMs (Fig. 3A).

We next compared the results of the COG assignment obtained by BLSOM (SomCOG) with those of the sequence homology search (HomCOG). Approximately 88%, 85% and 79% of the 4240 Sargasso HomCOG proteins were classified into COG categories same to those of BLAST on Tri11-, Di20- and Tetra6-BLSOMs, respectively. The highest level of congruity was found for Tri11-BLSOM. The reason why the 100% congruity and the complete overlap were not found (Fig. 3A) is thought to be related to the aforementioned finding that proteins belonging to different COGs, but having closely related functions, were associated with each other on BLSOMs. Levels of false-positives (i.e. assignment to unrelated COGs) were 6.3%, 6.4% and 7.1% on Tri11-, Di20- and Tetra6-BLSOMs, respectively, and the level was less than 0.2% for the overlapped zones of the two BLSOMs.

### 3.4. *Comparison of function predictions by BLSOM and BLAST*

In the above analysis, we focused on the Sargasso proteins assigned robustly to NcCOGs by BLAST (HomCOGs) in order to search for conditions giving high prediction accuracy. The most important contribution of the present alignment-free method should

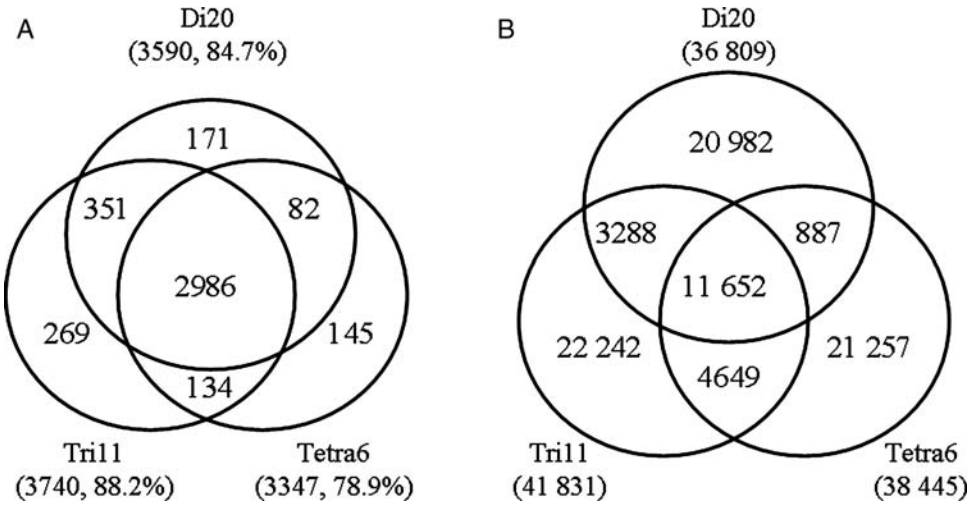


**Figure 3.** Venn diagrams representing COG predictions obtained by three BLSOMs. (A) Sargasso HomCOG proteins. The number and percentage in a parenthesis show the number of Sargasso proteins properly assigned to COG with each BLSOM and its percentage, respectively. (B) Sargasso proteins unassigned to COG with BLAST. The number in a parenthesis shows the number of Sargasso proteins assigned to COG with each BLSOM.

be to predict functions that cannot be predicted or can only be hypothesized ambiguously by the sequence homology search. To examine this potentiality, we used, as a test data set, the residual 93 598 Sargasso proteins that have been registered as ORFs in the International DNA Databanks but could not be robustly assigned to NcCOGs with BLAST. We mapped 200-amino acid segments derived from all 93 598 Sargasso proteins on the Di20-, Tri11- and Tetra6-BLSOMs constructed with NcCOG proteins, as described earlier. The overlap levels of the COG assignments for the residual 93 598 Sargasso proteins (Fig. 3B) were lower than those found for the 4240 Sargasso HomCOG proteins (Fig. 3A). This appears to fit the view that the residual Sargasso proteins presumably contain distant relatives of NcCOG proteins (e.g. paralogous genes) and are difficult to be assigned to a single NcCOG.

When the same prediction was obtained by different BLSOMs in Fig. 3B, the prediction would be meaningful, rather than accidental. This is because the three BLSOMs differed not only in the length of peptides (2, 3 and 4), but also in the categorization of amino acid residues (20, 11 and 6). A common SomCOG was assigned at least with two BLSOMs for 20 476 Sargasso proteins and was assigned with all three BLSOMs for 11 652 Sargasso proteins. To examine whether such predictions actually have biological meaning, BLAST search was conducted against NcCOG proteins, using each of the Sargasso COG proteins commonly assigned with three BLSOMs as a query. The identity and coverage levels found for the NcCOG protein with the lowest *e* value for each Sargasso query are plotted in Fig. 4A. Although a major portion of the data had more than 50% level of identity and more than 80% level of coverage, seven proteins had less than 30% identity, for which BLAST may not give convincing assignments. Next, we focused on Sargasso proteins assigned at least with two BLSOMs and conducted the above BLAST search. Data points with low identity and coverage levels increased significantly (Fig. 4B); for example, there were 66 proteins with less than 30% identity. In the final analysis, we focused on 63 752 Sargasso proteins that were assigned with any of the three BLSOMs and conducted the same BLAST search. The number of proteins with low levels of identity and coverage increased significantly, and there were 3097 proteins with less than 30% identity.

### 3.5. *Integrative assessment of function predictions by BLAST and BLSOM*

The sequence homology searches, such as BLAST and PSI-BLAST,[6] undoubtedly are essential tools for predicting protein function, but there has been accumulated a vast quantity of function-unknown proteins. One important contribution of BLSOM should be to provide an independent, separate prediction judgment, permitting us to get integrative assessment. For example, in the cases of Sargasso proteins with less than 60% levels of identity and coverage in the BLAST search, multiple COGs often became the candidates, and it was usually difficult to tell which of the candidates was the proper answer. When the BLSOM prediction fit to one of the candidates obtained by BLAST, this will become more reliable because the two methods are based on clearly distinct principles; sequence alignment-free and -dependent

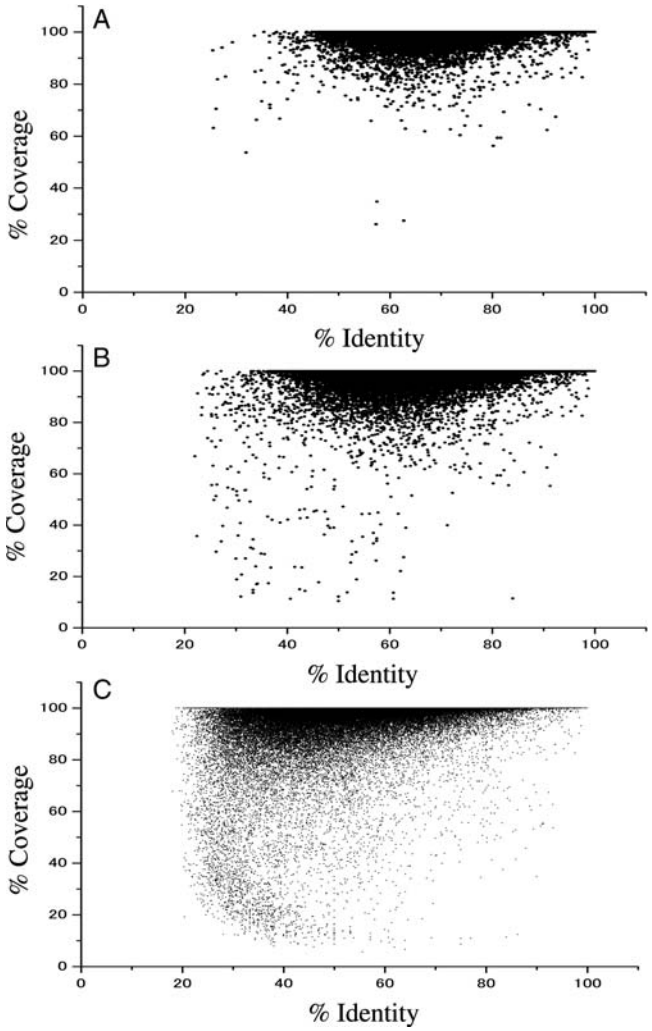


**Figure 4.** Identity and coverage levels found for the NcCOG protein with the lowest *e* value for each of Sargasso queries, which included HomCOG proteins. Sargasso proteins commonly assigned (A) with three BLSOMs, (B) with at least two BLSOMs and (C) with any BLSOM, respectively.

methods. In the cases of proteins with less than a 30% identity level (Fig. 4), convincing prediction will not be obtained with BLAST. For such twilight zone of alignments, various useful strategies have been developed to get reliable alignments.[1–8] BLSOM is a sequence alignment- and structural information-free method, and therefore, BLSOM prediction will become the novel, irreplaceable information.

The proof of functions, especially for novel proteins, can only be obtained by experimental approaches, and the purpose of any informatics prediction for such novel proteins is to provide the supporting-evidence/guideline for experimental studies. BLSOM will become a novel, irreplaceable strategy for function prediction of proteins for which only ambiguous or no prediction was obtained with sequence homology searches.

## 4. Discussion

Oligopeptides are component parts of a protein and involved in formation of both functional motifs and 3D structures. The BLSOM analyzing oligopeptide composition, therefore, appears to be positioned near with the functional motif searches, such as PRINTS and PROSITE,[29,30] which are powerful, indispensable methods for the function prediction of proteins. The functional motif search, however, does not adequately incorporate information pertaining to 3D structure formation. In the case of BLSOM, not only the oligopeptides for functional motifs, but also those contributing to 3D structure formation were included for predicting functions. Furthermore, because the functional motifs are obtained mostly from experimentally well-characterized proteins, the method might become less useful in the function prediction for the less-characterized proteins originating from poorly characterized genomes. The present unsupervised clustering algorism has the advantage that no advance knowledge about the target protein is required, and therefore, is thought to be appropriate for analyses of diverse novel proteins.

Sequence similarity and functional motif searches undoubtedly are essential, indispensable tools for predicting protein functions, but even after combining these powerful methods, a large number of function-unknown proteins remains. In order to complement these conventional methods and provide the additional new information helpful for integrative assessments, we developed the BLSOM method. The most important contribution of the present unsupervised, alignment-free method should be to predict functions of increasingly vast quantity of function-unknown proteins derived from less characterized genomes, such as those studied in the metagenomic approaches. For identifying functions of novel function-unknown proteins, a large-scale BLSOM that has analyzed all function-known proteins in databases must be constructed to extract their characteristics in advance. Various high-performance supercomputers are now available for biological studies, and the BLSOM developed here is suitable for actualizing high-performance parallel-computing with high-performance supercomputers such as the Earth Simulator 'ES'.[14–16] The present unsupervised, self-classifying strategy to find association of function-unknown proteins to function-known proteins on a large-scale BLSOM constructed with high-performance supercomputers should serve as a new and powerful tool in the post-genome era. In this study, we did not compare the performance of our approach with those of the structure-based methods, which would be one of our future works.

It is noteworthy that once the large-scale BLSOM is constructed using a high performance supercomputer and publicized, researchers can predict the function of individual proteins (e.g. those obtained from a newly sequenced genome) by mapping their proteins of interest on the large-scale BLSOM. This mapping is possible using a PC-level computer. The programs for the mapping, as well as the large-scale BLSOMs constructed by supercomputers, were available from our group (takaabe@nagahama-i-bio.ac.jp). This approach should add a new and powerful tool to the genomics and proteomics fields for the systematical and efficient prediction of functions of huge quantities of function-unknown proteins progressively accumulated in the International DNA Banks.

**Acknowledgements:** The computation was done with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

## Funding

This work was supported by Grant-in-Aid for Scientific Research (C) and for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

## References

1. Shi, J., Blundell, T.L. and Mizuguchi, K. 2001, FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties, *J. Mol. Biol.*, **310**, 243–57.
2. Andreeva, A., Howorth, D., Chandonia, J., et al. 2008, Data growth and its impact on the SCOP database: new developments, *Nucleic Acids Res.*, **36**, D419–25.
3. Cuff, A.L., Sillitoe, I., Lewis, T., et al. 2009, The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies, *Nucleic Acids Res.*, **37**, D310–4.
4. Todd, A.E., Orengo, C.A. and Thornton, J.M. 2001, Evolution of function in protein superfamilies, from a structural perspective, *J. Mol. Biol.*, **307**, 1113–43.
5. Lee, D., Redfern, O. and Orengo, C. 2007, Predicting protein function from sequence and structure, *Nat. Rev. Mol. Cell Biol.*, **8**, 995–1005.
6. Altschul, S.L., Madden, T.L., Schäffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
7. Rost, B. 1999, Twilight zone of protein sequence alignments, *Protein Eng.*, **12**, 85–94.
8. Chang, G.S., Hong, Y., Ko, K.D., et al. 2008, Phylogenetic profiles reveal evolutionary relationships within the "twilight zone" of sequence similarity, *Proc. Natl Acad. Sci. USA*, **105**, 1374–9.
9. Kohonen, T. 1990, The self-organizing map, *Proc. IEEE*, **78**, 1464–80.
10. Kohonen, T., Oja, E., Simula, O., Visa, A. and Kangas, J. 1996, Engineering applications of the self-organizing map, *Proc. IEEE*, **84**, 1358–84.
11. Ferran, E.A., Pflugfelder, B. and Ferrara, P. 1994, Self-organized neural maps of human protein sequences, *Protein Sci.*, **3**, 507–21.
12. Kanaya, S., Kinouchi, M., Abe, T., et al. 2001, Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli O157 genome, *Gene*, **276**, 89–99.
13. Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. 2003, Informatics for unveiling hidden genome signatures, *Genome Res.*, **13**, 693–702.
14. Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S. and Ikemura, T. 2005, Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples, *DNA Res.*, **12**, 281–90.
15. Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S. and Ikemura, T. 2006, A large-scale self-organizing map (SOM) unveils sequence characteristics of a wide range of eukaryote genomes, *Gene*, **365**, 27–34.
16. Abe, T., Sugawara, H., Kanaya, S. and Ikemura, T. 2006, Sequences from almost all prokaryotic, eukaryotic, and viral genomes available could be classified according to genomes on a large-scale self-organizing map constructed with the earth simulator, *J. Earth Simulator*, **6**, 17–23.
17. Uchiyama, T., Abe, T., Ikemura, T. and Watanabe, K. 2005, Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes, *Nat. Biotechnol.*, **23**, 88–93.
18. Hayashi, H., Abe, T., Sakamoto, M., et al. 2005, Direct cloning of genes encoding novel xylanases from human gut, *Can. J. Microbiol.*, **51**, 251–9.
19. Kosaka, T., Kato, S., Shimoyama, T., Ishii, S., Abe, T. and Watanabe, K. 2008, The genome of Pelotomaculum thermopropionicum reveals niche-associated evolution in anaerobic microbiota, *Genome Res.*, **18**, 442–8.
20. Abe, T., Sugawara, H., Kanaya, S. and Ikemura, T. 2006, A novel bioinformatics tool for phylogenetic classification of genomic sequence fragments derived from mixed genomes of environmental uncultured microbes, *Polar Biosci.*, **20**, 103–12.
21. Tatsusov, R.L., Koonin, E.V. and Lipman, D.J. 1997, A genomic perspective on protein families, *Science*, **278**, 631–7.
22. Lespinet, O., Wolf, Y.I., Koonin, E.V. and Aravind, L. 2002, The role of lineage-specific gene family expansion in the evolution of eukaryotes, *Genome Res.*, **12**, 1048–59.
23. Overbeek, R., Begley, T., Butler, R.M., et al. 2005, The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, *Nucleic Acids Res.*, **33**, 5691–702.

24. Koonin, E.V. and Aravind, L. 2008, Genomics of bacteria and Archaea: the emerging dynamic view of the prokaryotic world, *Nucleic Acids Res.*, **36**, 6688–719.
25. DeLong, E.F. 2002, Microbial population genomics and ecology, *Curr. Opin. Microbiol.*, **5**, 520–4.
26. Schloss, P.D. and Handelsman, J. 2003, Biotechnological prospects from metagenomics, *Curr. Opin. Biotechnol.*, **14**, 303–10.
27. Venter, J.C., Remington, K., Heidelberg, J.F., et al. 2004, Environmental genome shotgun sequencing of the Sargasso Sea, *Science*, **304**, 66–74.
28. Kosuge, T., Abe, T., Okido, T., et al. 2006, Exploration and grading of possible genes in 183 bacterial strains by a common fine protocol lead to new genes: Gene Trek in Prokaryote Space (GTPS), *DNA Res.*, **13**, 245–54.
29. Attwood, T. 2002, The PRINTS database: a resource for identification of protein families, *Brief Bioinform.*, **3**, 252–63.
30. Hulo, N., Bairoch, A., Bulliard, V., et al. 2007, The 20 years of PROSITE, *Nucleic Acids Res.*, **36**, D245–9.