Batch-Learning Self-Organizing Map for Predicting Functions of Poorly-Characterized Proteins Massively Accumulated

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

[1] Nagahama Institute of Bio-Science and Technology, Tamura-cho 1266, Nagahama-shi, Shiga-ken 526-0829, Japan
{takaabe,t_ikemura}@nagahama-i-bio.ac.jp
[2] Nara Institute of Science and Technology, Ikoma, Japan
skanaya@gtc.aist-nara.ac.jp

Abstract. As the result of the decoding of large numbers of genome sequences, numerous proteins whose functions cannot be identified by the homology search of amino acid sequences have accumulated and remain of no use to science and industry. Establishment of novel prediction methods for protein function is urgently needed. We previously developed Batch-Learning SOM (BL-SOM) for genome informatics; here, we developed BL-SOM to predict functions of proteins on the basis of similarity in oligopeptide composition of proteins. Oligopeptides are component parts of a protein and involved in formation of its functional motifs and structural parts. Concerning oligopeptide frequencies in 110,000 proteins classified into 2853 function-known COGs (clusters of orthologous groups), BL-SOM could faithfully reproduce the COG classifications, and therefore, proteins whose functions have been unidentified with homology searches could be related to function-known proteins. BL-SOM was applied to predict protein functions of large numbers of proteins obtained from metagenome analyses.

Keywords: batch-learning SOM, oligopeptide frequency, protein function, bioinformatics, high-performance supercomputer.

1 Introduction

Unculturable environmental microorganisms should contain a wide range of novel genes of scientific and industrial usefulness. Recently, a sequencing method for mixed genome samples directly extracted from environmental microorganism mixtures has become popular: metagenome analysis. A large portion of the environmental sequences thus obtained is registered in the International DNA Sequence Databanks (DDBJ/EMBL/GenBank) with almost no functional and phylogenetic annotation, and therefore, in the least useful manner. Homology searches for nucleotide and amino-acid sequences, such as BLAST, have become widely accepted as a basic bioinformatics tools not only for phylogenetic characterization of gene/protein sequences, but also for prediction of their biological functions when genomes and genomic segments are

J.C. Príncipe and R. Miikkulainen (Eds.): WSOM 2009, LNCS 5629, pp. 1–9, 2009.

decoded. Whereas the usefulness of the sequence homology search is apparent, it has became clear that homology searches can predict the protein function of only 50% or fewer of protein genes, when a novel genome or mixed genomes from environmental samples are decoded. In order to complement the sequence homology search, methods based on different principles are urgently required for predicting protein function.

Self-Organizing Map (SOM) is an unsupervised neural network algorithm developed by Kohonen and his colleagues [1-3], which provides an efficient and easy interpretation of the clustering of high-dimensional complex data using visualization on a two-dimensional plane. About 15 years ago, Ferran et al. [4] performed the pioneering and extensive SOM analysis of dipeptide composition in approximately 2000 human proteins stored in the SwissProt Database and reported clustering of the proteins according to both biological function and higher-order structure. Although this unsupervised learning method can be considered useful for predicting protein functions, the study was conducted long before decoding of genome sequences, and proteins of unknown function were rarely recognized at that time. Furthermore, because a long computation time was required for the SOM analysis of the dipeptide composition (400 dimensional vectorial data) even using high-performance computers at that time and because the final map was dependent on both the order of data input and the initial conditions, the conventional SOM method has rarely been used for prediction of protein function.

Previously, we developed a modified type SOM (batch-learning SOM: BL-SOM) for codon frequencies in gene sequences [5,6] and oligonucleotide frequencies in genome sequences [7-9] that depends on neither the order of data input nor the initial conditions. BL-SOM recognizes species-specific characteristics of codon or oligonucleotide frequencies in individual genomes, permitting clustering of genes or genomic fragments according to species without the need for species information during the BL-SOM learning. Various high-performance supercomputers are now available for biological studies, and the BL-SOM developed by our group is suitable for actualizing high-performance parallel-computing with high-performance supercomputers such as the Earth Simulator “ES” [10-12]. We previously used the BL-SOM for tetranucleotide frequencies for phylogenetic classification of genomic fragment sequences derived from mixed genomes of environmental microorganisms [13-16]. A large-scale phylogenetic classification was possible in a systematic way because genomic sequences were clustered (self-organized) according to species without species information or sequence alignment [7-9].

In the present report, we describe use of the BL-SOM method for prediction of protein function on the basis of similarity in composition of oligopeptides (di-, tri- and tetrapeptides in this study) of proteins. Oligopeptides are elementary components of a protein and are involved in the formation of functional motifs and structural organization of proteins. BL-SOM for oligopeptides may extract characteristics of oligopeptide composition, which actualize protein structure and function, and therefore, separate proteins according to their functions.

Sequences of approximately 8 million proteins are registered in the public databases, and about 500,000 proteins have been classified into approximately 5000 COGs (clusters of orthologous groups of proteins), which are the functional categories identified with bidirectional best-hit relationships between the completely sequenced genomes using the homology search of amino acid sequences [17]. The proteins

belonging to a single COG exhibit significant homology of amino acid sequences over the whole range of the proteins and most likely have the same function. Therefore, COG is undoubtedly a useful categorization of proteins according to function while the biological functions of a half of COGs have not yet been identified conclusively. In the present study, we initially focused on oligopeptide compositions in the 110,000 proteins classified into 2853 function-known COGs and prepared BL-SOMs under various conditions to search for conditions that would faithfully reproduce the COG classification. Then, we applied the BLSOM method to predict the functions of a large number of proteins obtained from metagenome analyses.

2 Methods

SOM implements nonlinear projection of multi-dimensional data onto a two-dimensional array of weight vectors, and this effectively preserves the topology of the high-dimensional data space [1-3]. We modified previously the conventional SOM for genome informatics on the basis of batch-learning SOM (BL-SOM) to make the learning process and resulting map independent of the order of data input [5-8]. The initial weight vectors were defined by PCA instead of random values, as described previously [6]. Weight vectors (w_{ij}) were arranged in the 2D lattice denoted by i (= 0, 1, . . . , I - 1) and j (= 0, 1, . . . , J - 1). I was set as 300 in Fig 1, and J was defined by the nearest integer greater than (s2/s1) × I; s1 and s2 were the standard deviations of the first and second principal components, respectively. Weight vectors (w_{ij}) were set and updated as described previously [5-8]. The 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 two days.

Amino acid sequences were obtained from the NCBI COG database (http://www.ncbi.nlm.nih.gov/COG/). Proteins shorter than 200 amino acids in length were

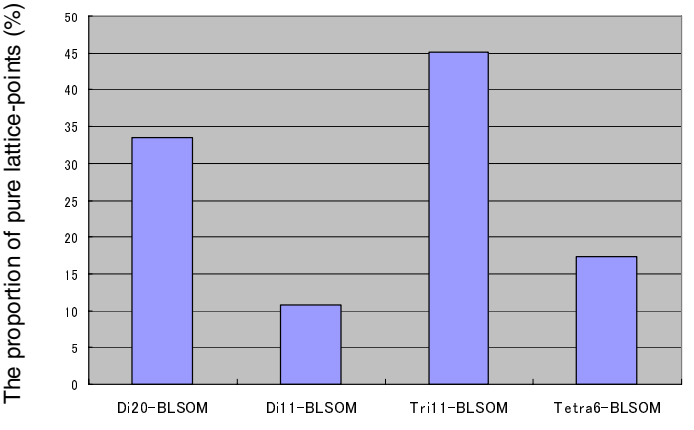

Fig. 1. The proportion of pure lattice-points for each condition

not included in the present study. We provided a window of 200 amino acids that is moved with a 50-amino acid step for proteins longer than 200 amino acids. To reduce the computation time, BL-SOM was constructed with tripeptide frequencies of the degenerate eleven groups of residues; {V, L, I}, {T, S}, {N, Q}, {E, D}, {K, R, H}, {Y, F, W}, {M}, {P}, {C}, {A} and {G}. BL-SOM 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}.

3 Results and Discussion

3.1 BL-SOMs Constructed with Proteins Belonging to COGs

For the test dataset to examine whether proteins are clustered (i.e., self-organized) according to function by BL-SOM, we chose proteins that had been classified into function known COGs by NCBI [17]. Using BL-SOM, dipeptide composition (20^2 = 400 dimensional vectorial data) was investigated in 110,000 proteins belonging to the 2853 function-known COGs. In addition to this BL-SOM for the dipeptide composition of 20 amino acids (abbreviated as Di20-BLSOM), the BL-SOM for the dipeptide or tripeptide composition were constructed after categorizing amino acids into 11 groups based on the similarity of their physicochemical properties, 11^2 (=121) or 11^3 (=1331) dimensional data (abbreviated as Di11- or Tri11-BLSOM, respectively). BL-SOM was also constructed for the tetrapeptide composition after categorization into 6 groups, 6^4 (=1296) dimensional data (abbreviated as Tetra6-BLSOM). These four different BL-SOM conditions were examined to establish which gave the best accuracy and to what degree similar results were obtained among the four conditions. It should be noted that BL-SOMs 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 (14641-dimensional data), was difficult in the present study because of the limitations of ES resources available to our group.

In order to introduce a method that is less dependent on the sequence length of proteins, 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 BL-SOM was constructed for these overlapped 200-amino acid sequences (approximately 500,000 sequences in total). Introduction of a window with a shifting step enabled us to analyze both multifunctional multidomain proteins primarily originating from the fusion of distinct proteins during evolution and smaller proteins, collectively. The 200-amino acid window was tentatively chosen in the present study because sizes of fictional domains of a wide range of proteins range from 100 to 300 amino acids.

One important issue of the present method is at what level each lattice-point on a BL-SOM contains 200-amino acid fragments derived from a single COG. The number of the function-known COG categories is 2835, and the size of the BL-SOM was chosen so as to provide approximately eight fragments per lattice-point on average. If fragments were randomly chosen, the probability that all fragments associated with one lattice-point were derived from a single COG by chance should be extremely low, e.g. $(1/2853)^8 = 2.3\times10^{-28}$, while ensuring that this value depends on

the total number of fragments derived from proteins belonging to the respective COG. We designate here the lattice-point that contained fragments derived only from a single COG category as a "pure lattice-point".

We compared the occurrence level of pure lattice-points among four different BL-SOMs. Although no COG information was given during BL-SOM learning, a high percentage of pure lattice-points (i.e., correct self-organization of sequences according to the COG category) was obtained (Fig. 1), despite the fact that the occurrence probability of a pure lattice-point as an accidental event is extremely low. The highest occurrence level of pure lattice-points was observed on the Tri11-BLSOM; approximately 45% of lattice-points on the Tri11-BLSOM contained sequences derived from only a single COG (Fig. 1). To graphically show the difference among these BL-SOMs, pure lattice-points were colored in red (Fig. 2A-C). The finding that the COG clustering (self-organization) with high accuracy was achieved indicates BL-SOM to be a powerful tool for function prediction of function-unknown proteins.

In Fig. 3, the number of sequences at each pure lattice-point on the Tri11-BLSOM was shown with the height of the vertical bar with a color representing each of the

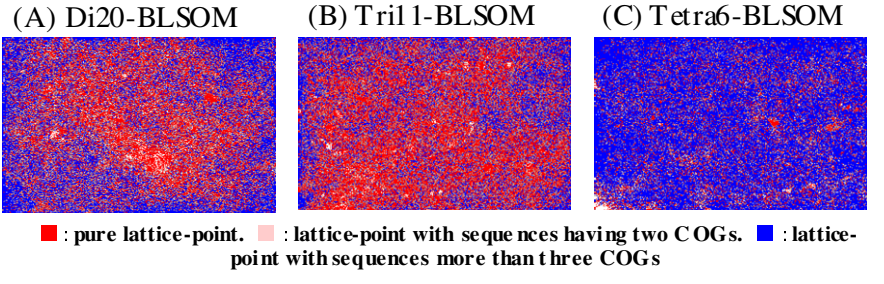

Fig. 2. The distribution of pure lattice-points colored in red

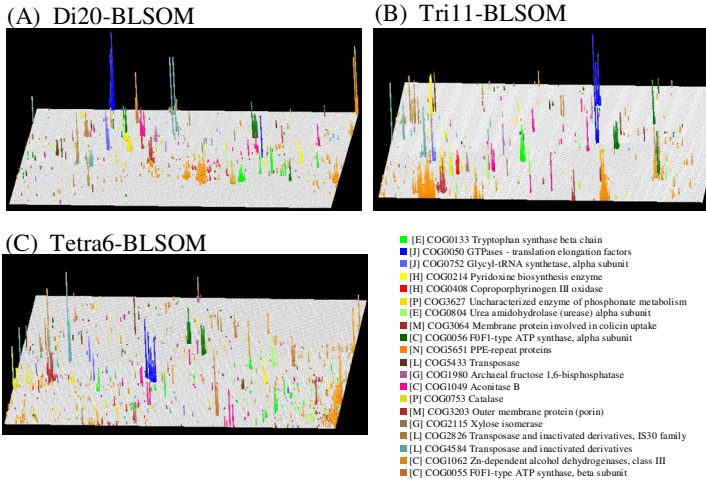

Fig. 3. Clustering of protein sequences according to COG (20 samples)

20 COG examples. Not only for these 20 examples, but also for a large portion of COG categories, sequences belonging to a single COG were localized in the neighboring points, resulting in a high peak composed of neighboring, high bars. In Fig. 3, a few high peaks with the same color located far apart from each other are also observed. Detailed inspection showed that these detached high peaks were mostly due to the different 200-amino acid segments (e.g., anterior and posterior portions) derived from one protein, which have distinct oligopeptide compositions and possibly represented distinct structural and functional domains of the respective protein. This type of major but distinct peaks appears to be informative for the prediction of functions of multifunctional multidomain proteins.

3.2 Function Prediction of Proteins Obtained from Metagenome Analyses

Most environmental microorganisms cannot be cultured under laboratory conditions. Genomes of the unculturable microorganisms have remained mostly uncharacterized but are believed to contain a wide range of novel protein genes of scientific and industrial usefulness. Metagenomic approaches that decode the sequences of the mixed genomes of uncultured environmental microbes [18-20] have been developed recently for finding a wide variety of novel and industrially useful genes. Venter et al. [21] applied large-scale metagenome sequencing to mixed genomes collected from the Sargasso Sea near Bermuda and deposited a large number of sequence fragments in the International DNA Databanks.

The most important contribution of the present alignment-free and unsupervised clustering method, BLSOM, should be the prediction of the functions of an increasingly large number of function-unknown proteins derived from the less characterized genomes, such as those studied in the metagenomic approaches. To test the feasibility of BL-SOM for function prediction of environmental sequences, we searched in advance the Sargasso proteins that showed significant global homology with the NCBI COG proteins by using the conventional sequence homology search. Based on a criterion that in 80% or more of the region, 80% or more identity of the amino acid sequence was observed, 3924 Sargasso proteins (> 200 amino acids) could be related to NCBI COG categories (designated Sargasso COG sequences). Then, we mapped the 200-amino acid fragments derived from these Sargasso COG proteins onto Di20- and Tri11-BLSOMs, which were previously constructed with NCBI COG sequences in Figs. 1 and 2. For each lattice point on which Sargasso COG fragments were mapped, the most abundant NCBI COG sequences were identified, and the mapped Sargasso segments were tentatively assumed to belong to this most abundant NCBI COG category. After summing up these tentative COGs for each Sargasso protein, individual Sargasso proteins were finally classified into one NCBI COG category, if more than 60% of the 200-amino acid fragments derived from one Sargasso protein gave the same COG category. By mapping on Tri11-, Di20- or Tet6-BLSOM, 87.5, 86.8 or 79.0% of the 3924 Sargasso COG proteins showed the COG category identical to that had been identified by the sequence homology search in advance. The highest identity level was found on Tri11-BLSOM. In Fig. 4, the number of Sargasso fragments thus classified into COGs on Di20-, Tri11- and Tetra6-BLSOM was shown by the height of the vertical bar. In the next step, when the false prediction for the Sargasso COG proteins was checked in detail, the pairs of real and

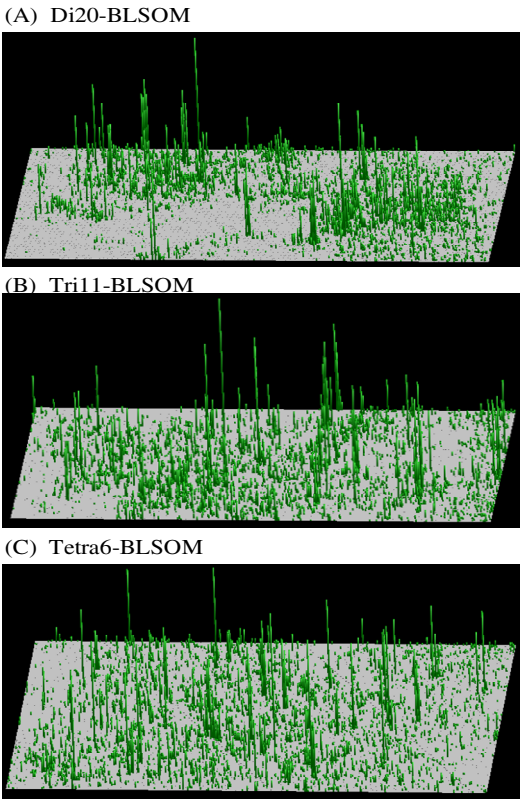

Fig. 4. Mapping of the Sargasso COG fragments on Di20- (A), Tri11- (B) and Tetra6- (C) BLSOM. The height of the vertical bar shows the number of fragments.

falsely-assigned COGs corresponded to those that have the functions closely related with each other, such as those with paralogous relationships. According to the definition of COG (clusters of orthologous groups of proteins), paralogous gene proteins should belong to different COGs in spite of the similarity of functions. COG categorization appears be too strict to be used for function predictions of a wide variety of proteins.

In the final analysis, we mapped the residual Sargasso proteins, which could not be classified into NCBI COGs using the sequence homology search, onto Di20-, Tri11- and Tetra6-BLSOMs. Approximately 15% of the Sargasso proteins (i.e., approximately 90,000 proteins) were associated with an NCBI COG category. For Sargasso proteins for which the consistency of the predicted function is obtained by separate analyses of di-, tri- and tetrapeptide frequencies, the reliability of the prediction should be very high. We plan to publicize the results of the assignments obtained concordantly with three BLSOM conditions (Tri11-, Di20- and Tetra6-BLSOMs).

To identify functions of a large number of function-unknown proteins accumulated in databases systematically, we have to construct a large scale-BL-SOM in advance

that analyzes all function-known proteins available in databases utilizing a high-performance supercomputer such as ES [22,23]. This approach should serve as a new and powerful strategy to predict functions of a large number of novel proteins collectively, systematically and efficiently. The BLSOM data obtained by high-performance supercomputers are unique in fields of genomics and proteomics and provide a new guideline for research groups, including those in industry, for the study of function identification of novel genes through experiment.

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

References

1. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59–69 (1982)
2. Kohonen, T.: The self-organizing map. Proc. IEEE 78, 1464–1480 (1990)
3. Kohonen, T., Oja, E., Simula, O., Visa, A., Kangas, J.: Engineering applications of the self-organizing map. Proc. IEEE 84, 1358–1384 (1996)
4. Ferran, E.A., Pflugfelder, B., Ferrara, P.: Self-organized neural maps of human protein sequences. Protein Sci. 3, 507–521 (1994)
5. Kanaya, S., Kudo, Y., Abe, T., Okazaki, T., Carlos, D.C., Ikemura, T.: Gene classification by self-organization mapping of codon usage in bacteria with completely sequenced genome. Genome Inform. 9, 369–371 (1998)
6. Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H., Ikemura, T.: 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 (2001)
7. Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., Ikemura, T.: A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: Self-organizing map of oligonucleotide frequency. Genome Inform. 13, 12–20 (2002)
8. Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T., Ikemura, T.: Informatics for unveiling hidden genome signatures. Genome Res. 13, 693–702 (2003)
9. Abe, T., Kozuki, T., Kosaka, Y., Fukushima, S., Nakagawa, S., Ikemura, T.: Self-organizing map reveals sequence characteristics of 90 prokaryotic and eukaryotic genomes on a single map. In: WSOM 2003, pp. 95–100 (2003)
10. Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S., Matsuura, Y., Tokutaka, H., Ikemura, T.: A large-scale Self-Organizing Map (SOM) constructed with the Earth Simulator unveils sequence characteristics of a wide range of eukaryotic genomes. In: WSOM 2005, pp. 187–194 (2005)
11. Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S., Ikemura, T.: A large-scale Self-Organizing Map (SOM) unveils sequence characteristics of a wide range of eukaryote genomes. Gene. 365, 27–34 (2006)
12. Abe, T., Sugawara, H., Kanaya, S., Ikemura, T.: 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 (2006)

13. Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S., Ikemura, T.: Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Res. 12, 281–290 (2005)
14. Hayashi, H., Abe, T., Sakamoto, M., et al.: Direct cloning of genes encoding novel xylanases from human gut. Can. J. Microbiol. 51, 251–259 (2005)
15. Uchiyama, T., Abe, T., Ikemura, T., Watanabe, K.: Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nature Biotech. 23, 88–93 (2005)
16. Abe, T., Sugawara, H., Kanaya, S., Ikemura, T.: A novel bioinformatics tool for phylogenetic classification of genomic sequence fragments derived from mixed genomes of environmental uncultured microbes. Polar Bioscience 20, 103–112 (2006)
17. Tatsusov, R.L., Koonin, E.V., Lipman, D.J.: A genomic perspective on protein families. Science 278, 631–637 (1997)
18. Amann, R.I., Ludwig, W., Schleifer, K.H.: Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169 (1995)
19. Hugenholtz, P., Pace, N.R.: Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. Trends Biotechnol. 14, 190–197 (1996)
20. Rondon, M.R., August, P.R., Bettermann, A.D., et al.: Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66, 2541–2547 (2000)
21. Venter, J.C., et al.: Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74 (2004)
22. Abe, T., Ikemura, T.: A large-scale batch-learning Self-Organizing Maps for function prediction of poorly characterized proteins progressively accumulating in sequence databases. Annual Report of the Earth Simulator, April 2006 - March 2007, pp. 247–251 (2007)
23. Abe, T., Ikemura, T.: A large-scale genomics and proteomics analyses conducted by the Earth Simulator. Annual Report of the Earth Simulator, April 2007 - March 2008, pp. 245–249 (2008)