An artificial intelligence approach fit for tRNA gene
studies in the era of big sequence data

Yuki Iwasaki1‡, Takashi Abe2*, Kennosuke Wada1, Yoshiko Wada1 and Toshimichi Ikemura1†

1Department of Bioscience, Nagahama Institute of Bio-Science and Technology, 1266 Tamura-Cho, Nagahama, Shiga 526-0829, Japan
2Department of Information Engineering, Faculty of Engineering, Niigata University, 2-8050, Ikarashi, Nishi-ku, Niigata 950-2181, Japan

(Received 15 December 2016, accepted 29 January 2017; J-STAGE Advance published date: 24 March 2017)

Unsupervised data mining capable of extracting a wide range of knowledge from big data without prior knowledge or particular models is a timely application in the era of big sequence data accumulation in genome research. By handling oligonucleotide compositions as high-dimensional data, we have previously modified the conventional self-organizing map (SOM) for genome informatics and established BLSOM, which can analyze more than ten million sequences simultaneously. Here, we develop BLSOM specialized for tRNA genes (tDNAs) that can cluster (self-organize) more than one million microbial tDNAs according to their cognate amino acid solely depending on tetra- and pentanucleotide compositions. This unsupervised clustering can reveal combinatorial oligonucleotide motifs that are responsible for the amino acid-dependent clustering, as well as other functionally and structurally important consensus motifs, which have been evolutionarily conserved. BLSOM is also useful for identifying tDNAs as phylogenetic markers for special phylogenotypes. When we constructed BLSOM with ‘species-unknown’ tDNAs from metagenomic sequences plus ‘species-known’ microbial tDNAs, a large portion of metagenomic tDNAs self-organized with species-known tDNAs, yielding information on microbial communities in environmental samples. BLSOM can also enhance accuracy in the tDNA database obtained from big sequence data. This unsupervised data mining should become important for studying numerous functionally unclear RNAs obtained from a wide range of organisms.

Key words: tRNA gene, metagenome, oligonucleotide composition, big data, phylogenetic marker

INTRODUCTION

Compilation of tRNA sequences and genes was originally established by Sprinzl and coworkers (Sprinzl et al., 1978; Sprinzl and Vassolenko, 2005), and has been updated (tRNAdb; http://tRNAdb.bioinf.uni-leipzig.de/) (Jühling et al., 2009). Using tRNAscan-SE, the Genomic tRNA Database (GtRNAdb; http://lowelab.ucsc.edu/GtRNAdb/) has been constructed for complete and near-complete genomes (Chan and Lowe, 2009). In addition, the genomic organization of eukaryotic tRNAs has been extensively studied, and shows complex lineage-specific variability (Bermudez-Santana et al., 2010). For both completely and partially sequenced genomes, as well as vast numbers of metagenomic sequences from a wide variety of environmental and clinical samples, we have constructed and updated a large-scale database of tRNA genes (tDNAs) called tRNADB-CE (http://trna.ie.niigata-u.ac.jp) (Abe et al., 2011). Metagenomic sequences have attracted broad scientific and industrial interest, and even short sequences obtained with new-generation sequencers (e.g., Sequence Read Achieve in NCBI, http://www.ncbi.nlm.nih.gov/Traces/sra/) contain numerous full-length tDNAs because of their short length. The number of tDNAs compiled in tRNADB-CE is already large (1.7 million genes) and will undoubtedly increase rapidly in the future. For efficient knowledge discovery from such big data, new tools are important for promoting promising research on genes for tRNAs and other RNAs including a wide variety of function-unclear RNAs.

In the current era of big sequence data obtained from high-throughput DNA sequencers, it is important to
establish an unsupervised data mining method capable of extracting a wide range of knowledge without prior knowledge, hypotheses, or particular models from numerous genomic sequences (e.g., tDNA sequences) covering a wide range of species, for which experimental studies other than DNA sequencing are often lacking. Various unsupervised data mining methods, such as K-means clustering and Fuzzy Art (Forgy, 1965; Carpenter et al., 1991; Hastie et al., 2009), have been developed; and we have previously developed an unsupervised clustering method, BLSOM (batch-learning self-organizing map) (Kanaya et al., 2001; Abe et al., 2003, 2005; Kikuchi et al., 2015), which can analyze more than ten million genomic sequences simultaneously and allows acquisition of a wide range of knowledge from big sequence data. For example, BLSOM with oligonucleotide (e.g., tetrancleotide) composition can cluster genomic sequence fragments (e.g., 1-kb sequences) according to phylotype, and has thus succeeded in phylogenetic classification of a large number of metagenomic sequences (Uchiyama et al., 2005; Uehara et al., 2011; Nakao et al., 2013).

Oligonucleotides, such as penta- and hexanucleotides, often represent motif sequences that are responsible for sequence-specific protein binding such as transcription factor binding, and their occurrences should differ from those expected from mononucleotide composition in each genome and among genomic portions within one genome. An analysis of the human genome with pentanucleotide BLSOM has unexpectedly found evident enrichment of many kinds of transcription factor-binding motifs in pericentric heterochromatin regions (Iwasaki et al., 2013), showing that BLSOM effectively detects characteristic, combinatorial occurrences of functional-motif oligonucleotides in genomic sequences with no prior knowledge.

BLSOM is suitable for actualizing high-performance parallel computing, and thus for the analysis of high-dimensional big data. Here, we have tested its usefulness for data mining from a large number of tDNA sequences. We found that BLSOM for tDNAs can reveal combinatorial oligonucleotide motifs that are responsible for their amino acid-dependent clustering, and BLSOM for species-unknown tDNAs obtained from metagenomic sequences plus species-known microbial tDNAs can provide information on microbial communities in environmental samples.

MATERIALS AND METHODS

**tDNA sequences** To enhance the completeness and accuracy of tDNAs compiled in tRNADB-CE, three computer programs, tRNAscan-SE (Lowe and Eddy, 1997), ARAGORN (Laslett and Canback, 2004) and tRNAfinder (Kinouchi and Kurokawa, 2006) were used in combination, since their algorithms partially differ and rendered somewhat different results. tDNAs found concordantly by the three programs were stored in tRNADB-CE and discordant cases among programs were manually checked by experts in tRNA experimental fields (Abe et al., 2011). The present study has constructed BLSOMs for tri-, tetra-, and pentanucleotide compositions in tDNAs in tRNADB-CE (Version 7.0; last update, 2014/01/25). Since a portion of tDNAs lack the terminal CCA sequence, and one purpose of this study is to establish a strategy for using tDNAs as phylogenetic markers, the CCA terminus sequence has been excluded from BLSOM analyses in the present study. However, inclusion of the CCA sequence for BLSOM analyses of tDNAs may enhance amino acid-dependent clustering for tDNAs containing the CCA sequence; the CCA terminus is abundant in RNA-seq data (Findeiss et al., 2011) and its importance of the CCA terminus in the genomic tag has been proposed (Weiner and Maizels, 1987). Therefore, the combinatorial use of CCA-plus and -minus analyses may provide new additional information.

**BLSOM** SOM (self-organizing map) is an unsupervised clustering algorithm that nonlinearly maps high-dimensional vectorial data onto a two-dimensional array of lattice points; i.e., a flexible net that is spread into the multi-dimensional “data cloud” (Kohonen, 1982; Kohonen et al., 1996). We previously modified the conventional SOM for genome informatics to make the learning process and resulting map independent of the order of data input on the basis of batch-learning SOM, or BLSOM (Kanaya et al., 2001). The initial vectors were defined by principal component analysis (PCA) instead of random values.

The frequency of each pentanucleotide obtained from vectorial data representing each lattice point on Penta in Fig. 1A is calculated and normalized with the level expected from the mononucleotide composition, calculated from vectorial data representing the lattice point. The observed/expected ratio is illustrated in red (overrepresented), blue (underrepresented) and white (moderately represented) according to Abe et al. (2003) (Fig. 1D). Since there are $1,024 = 4^5$ different pentanucleotides, the occurrence of one pentanucleotide in one tDNA is primarily 0 or 1. Thus, red and blue primarily show the presence and absence of each pentanucleotide.

**Parts-BLSOM** Computer programs used to find tDNAs can divide each tDNA sequence into the following structural parts: 5’ side of acceptor stem, D-arm, anticodon arm, variable-arm (V-arm), T-arm, 3’ side of acceptor stem and CCA terminus. BLSOM in which oligonucleotides found in different structural parts are differentially counted is designated Parts-BLSOM (for the correspondence with the secondary cloverleaf structure, see Supplementary Fig. S1). Since the CCA terminus sequence has been
excluded from the present BLSOM analysis, PartsTri and PartsTetra treat $384 (= 64 \times 6)$ and $1,536 (= 256 \times 6)$ variables, respectively. PartsPenta was not included because of the V-arm’s short length for several amino acids; e.g., shorter than 5 nt for Glu, Cys, and Gly. BLSOM programs can be obtained from our web site (http://bioinfo.ie.niigata-u.ac.jp/BLSOM). Distances of weight vectors between neighboring lattice points on BLSOM can be visualized as black levels with a U-matrix method (Ultsch, 1993), and this provides information about similarity of oligonucleotide composition in local areas on BLSOM (Iwasaki et al., 2013).

RESULTS

Oligonucleotide BLSOMs for bacterial tDNAs Each tRNA has characteristic combinatorial occurrences of various motif oligonucleotides that are required to fulfill its function (e.g., binding to proper enzymes and rRNAs) and to form its structure (L-shaped form). To examine the usefulness of BLSOM for efficient knowledge discovery from massive numbers of tDNAs, we conducted BLSOMs for tri-, tetra- and pentanucleotide compositions in approximately 0.4 million tDNAs from more than 7,000 bacterial genomes that are categorized as “Reliable tRNAs” in tRNADB-CE (Tri, Tetra and Penta in Fig. 1A); archaean and fungal tDNAs will be analyzed later. Lattice points containing multiple tDNAs, but not one or a few tDNAs, are detectable. A satellite single bar (arrowed) located between two major Gly territories is composed of 34 Gly-GCC tDNAs belonging to six *Chlamydophila* species: *C. caviae*, *C. felis*, *C. muridarum*, *C. pneumoniae* and *C. trachomatis*. (D) Examples of diagnostic pentanucleotides responsible for amino acid-dependent clustering on Penta in Fig. 1A. The occurrence of each pentanucleotide for each lattice point was calculated and normalized with occurrence expected from the mononucleotide composition for the respective lattice point (Abe et al., 2005). This observed/expected ratio is indicated in color: red (overrepresented), blue (underrepresented), blank (intermediate). (E) An example of pentanucleotides, TTCGA, observed for most bacterial tDNAs. Lattice points are marked as described in B.
obtained with no information other than oligonucleotide composition. Thus, BLSOM should be able to pick out characteristic combinations of motif sequences required for proper recognition by various enzymes, such as aminoacyl-tRNA synthetase (aaRS), and of sequences supporting proper L-shaped structures.

Figure 1B marks lattice points containing tDNAs of individual amino acids separately on Penta (for other amino acids, including selenocysteine, see Supplementary Fig. S2). tDNAs for one amino acid form one or a few major territories, as well as many tiny satellite-type spots. It should be noted that the major territory located at the bottom of the Met panel is composed solely of initiator Met tDNAs, but the upper major territory is composed of both elongator Met tDNAs and Ile tDNAs containing the anticodon CAT, which is enzymatically converted to read the ATA codon. When considering the biological significance of minor territories and tiny satellites, the number of tDNAs in each lattice point is important. Thus, the vertical bars in Fig. 1C present the number of Gly tDNAs (for other amino acids, see Supplementary Fig. S2). Lattice points in two major Gly territories apparently contain many tDNAs, and one tiny satellite located away from major territories also has multiple tDNAs (arrowed in Fig. 1C), which represent 34 Gly-GCC tDNAs with two base differences and derived from six Chlamydophila species listed in the figure legend. Similar satellite-type peaks are observed for other amino acids (Supplementary Fig. S2) and represent isoacceptors of various species primarily belonging to one phylogenetic family, which often differ in sequence by only a few bases; examples of multiple alignment of such sequences are presented along with their phylotypes in Supplementary Fig. S3: Arg-GCG for four Borrelia species, and Asn-GTT for six Thermotogae species. These types of noncanonical tDNAs are candidates for molecular phylogenetic markers representing a specific phylotype.

BLSOM clusters tDNAs according to amino acid, solely depending on oligonucleotide composition, and visualizes major, minor and noncanonical rare tDNAs. BLSOM is an unsupervised clustering algorithm and allows us to explore causative factors responsible for the amino acid-dependent clustering (self-organization) and to compare causative factors pointed out by BLSOM with known molecular mechanisms experimentally proven for a limited number of model organisms, such as the mechanisms reviewed by Marck and Grosjean (2002). Importantly, we can address the following well-timed questions in the era of big sequence data accumulation: to what range of phylotypes can a certain known molecular mechanism (e.g., sequence motifs recognized by an aaRS) be applied, and what types of alternative mechanisms can be expected for other phylotypes? Since experimental studies are limited for most sequenced genomes, this in silico characterization has become increasingly important, and BLSOM has powerful visualization functions that are useful for addressing these questions.

**Visual** **ization of combinatorial occurrences of functionally important oligonucleotides** Functionally and structurally important sequences in tRNAs have been stably maintained throughout evolution. Therefore, a wide range of species has the very closely related motifs, while sequences outside the motifs have diverged significantly. For example, to identify cognate isoacceptors from a pool of tRNAs sharing a similar L-shaped structure, aaRS recognizes a relatively small number of nucleotides as RNA code (Schimmel et al., 1993) and identity elements (Normanly and Abelson, 1989; Ibba and Söll, 2000; Ardell, 2010). Oligonucleotide sequences maintained stably in a large number of isoacceptors from a wide range of bacteria should be responsible and diagnostic for their amino acid-dependent clustering (self-organization) on BLSOM. This prediction has been proven by using the BLSOM capability to visualize diagnostic oligonucleotides contributing to self-organization, as explained in Materials and Methods. Red and blue in Fig. 1D and E show the presence and absence of the respective pentanucleotide on Penta in Fig. 1A. Transitions between red and blue for various pentanucleotides often coincide with borders between territories of different amino acids, and Fig. 1D shows three examples of

---

**Table 1. Percentages of tDNAs located at colored pure lattice points on BLSOM**

<table>
<thead>
<tr>
<th></th>
<th>Tri</th>
<th>Tetra</th>
<th>Penta</th>
<th>Parts/Tri</th>
<th>Parts/Tetra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>90.0</td>
<td>96.5</td>
<td>95.8</td>
<td>99.1</td>
<td>98.4</td>
</tr>
<tr>
<td>Arg</td>
<td>76.9</td>
<td>93.0</td>
<td>95.6</td>
<td>96.9</td>
<td>97.1</td>
</tr>
<tr>
<td>Asn</td>
<td>83.1</td>
<td>93.8</td>
<td>96.5</td>
<td>98.0</td>
<td>99.4</td>
</tr>
<tr>
<td>Asp</td>
<td>94.3</td>
<td>98.5</td>
<td>98.8</td>
<td>98.6</td>
<td>99.0</td>
</tr>
<tr>
<td>Cys</td>
<td>81.3</td>
<td>94.5</td>
<td>95.1</td>
<td>96.7</td>
<td>98.2</td>
</tr>
<tr>
<td>Gln</td>
<td>86.3</td>
<td>94.6</td>
<td>94.5</td>
<td>91.2</td>
<td>96.7</td>
</tr>
<tr>
<td>Glu</td>
<td>94.5</td>
<td>97.9</td>
<td>97.5</td>
<td>98.3</td>
<td>98.3</td>
</tr>
<tr>
<td>Gly</td>
<td>89.2</td>
<td>96.7</td>
<td>98.1</td>
<td>98.3</td>
<td>98.8</td>
</tr>
<tr>
<td>His</td>
<td>78.8</td>
<td>93.9</td>
<td>91.0</td>
<td>95.5</td>
<td>95.7</td>
</tr>
<tr>
<td>Ile</td>
<td>88.8</td>
<td>97.5</td>
<td>93.8</td>
<td>95.1</td>
<td>97.5</td>
</tr>
<tr>
<td>Leu</td>
<td>88.4</td>
<td>97.3</td>
<td>98.4</td>
<td>98.7</td>
<td>98.3</td>
</tr>
<tr>
<td>Lys</td>
<td>77.3</td>
<td>92.9</td>
<td>95.9</td>
<td>95.8</td>
<td>97.1</td>
</tr>
<tr>
<td>Met</td>
<td>86.4</td>
<td>95.8</td>
<td>95.6</td>
<td>96.6</td>
<td>98.1</td>
</tr>
<tr>
<td>Phe</td>
<td>83.3</td>
<td>98.3</td>
<td>98.3</td>
<td>99.3</td>
<td>98.9</td>
</tr>
<tr>
<td>Pro</td>
<td>81.8</td>
<td>93.5</td>
<td>95.8</td>
<td>97.9</td>
<td>98.9</td>
</tr>
<tr>
<td>Ser</td>
<td>92.5</td>
<td>95.2</td>
<td>96.3</td>
<td>98.3</td>
<td>98.7</td>
</tr>
<tr>
<td>Thr</td>
<td>78.2</td>
<td>91.8</td>
<td>90.1</td>
<td>96.9</td>
<td>97.6</td>
</tr>
<tr>
<td>Trp</td>
<td>82.7</td>
<td>98.6</td>
<td>91.7</td>
<td>90.8</td>
<td>91.9</td>
</tr>
<tr>
<td>Tyr</td>
<td>90.7</td>
<td>97.9</td>
<td>98.8</td>
<td>98.8</td>
<td>98.8</td>
</tr>
<tr>
<td>Val</td>
<td>86.4</td>
<td>94.9</td>
<td>94.1</td>
<td>97.4</td>
<td>97.2</td>
</tr>
</tbody>
</table>
pentanucleotides observed mainly in major territories of one amino acid and most likely related to identity elements. The major red zone of ATAGA corresponds to the major territory of Arg in Fig. 1B; this pentanucleotide exists in the D-arm in a major portion of bacterial Arg tDNAs. The major red zones of GATAA and CATAA correspond to major territories of Ile and Met tDNAs seen in Fig. 1B; these pentanucleotides exist in the anticodon arm in these isoacceptors. In fact, anticodon arm and D(dihydroU)-arms have been reported to contain highly significant identity elements; for example, Ardell (2010) has systematically compiled identity determinants in Proteobacteria. While a major type of identity element for one amino acid has been well conserved in sequence throughout evolution, mechanisms for aaRS to recognize cognate tRNAs seem to have diverged to some extent, even among bacterial species (Marck and Grosjean, 2002; Ardell, 2010). tDNAs with minor, noncanonical identity elements can be detected by identifying tDNAs located outside red zones of the pentanucleotide representing a canonical identity element. Such tDNAs should become phylogenetic markers for a restricted phylogenetic lineage, and real examples will be mentioned later.

We next examine canonical-type oligonucleotides present in a large portion of bacterial tDNAs. For example, the functionally important and well conserved heptanucleotide GGTTCGA in the T-arm (abbreviated to T-arm) is observed for a large majority of bacterial tDNAs (Marck and Grosjean, 2002; Ardell, 2010) and, therefore, the three constituent pentanucleotides of this heptanucleotide are observed in a major portion of lattice points on Penta, as colored in red in Fig. 1E (TTCGA): the other two pentanucleotides give similar (but not identical) results. Small blue areas contain tDNAs that lack the consensus sequence. Actually, the aforementioned *Chlamydomphila* Gly-GCC tDNAs (arrowed in Fig. 1C) differ from the canonical heptanucleotide in the T-arm at two bases and thus are located in a blue zone (arrowed in Fig. 1E) for all three constituent pentanucleotides. This is one reason that *Chlamydomphila* Gly-GCC tDNAs form a satellite peak outside the Gly major territories in Fig. 1C, and shows that tDNAs with noncanonical sequences in the T-arm are strong candidates for phylogenetic markers. The occurrence levels of the three pentanucleotides for each amino acid are presented in Supplementary Fig. S4. These pentanucleotides are observed in more than 70% of bacterial tDNAs of Ala, Arg, Asn, Ile, Lys, Met, Phe, Thr and Val, but almost no tDNAs of Cys, Gln, Glu, Leu, Ser and Try. For Asp, Gly, His, Pro and Trp, a portion of tDNAs have these pentanucleotides. Some tDNAs belonging to the last five amino acids may become candidates for phylogenetic markers after clarification of their existence/non-existence according to the phylogenetic group.

**BLSOM specialized for tDNA research: Parts-BLSOM** Tri, Tetra, and Penta in Fig. 1 can cluster amino acid-specific tDNAs with no information other than oligonucleotide composition and predict functionally and/or structurally important motifs, as exemplified in Fig. 1D and E. This is a favorable feature of unsupervised data mining. These BLSOMs, however, do not take into consideration the fundamental characteristic of tRNAs that functionally and structurally important motifs exist in specific parts of the molecule. Thus, an oligonucleotide that happens to have the same sequence as a functional motif (e.g., an identity element) but is located outside the functional site cannot be distinguished from the real functional element. Actually, even outside the characteristic territories for one amino acid, there are small red zones for the pentanucleotide related to the identity element of the amino acid (Fig. 1D), and these pentanucleotides have often been found outside the identity element in noncognate tRNAs. We next add the following information about tRNA molecules other than oligonucleotide composition and examine the usefulness of this change.

Here, we construct a new BLSOM, in which oligonucleotides found in six different structural parts are differentially counted. Tri- and tetranucleotide BLSOMs with 384 (= 64 × 6) and 1,536 (= 256 × 6) variables have been constructed for bacterial tDNAs, as described in Materials and Methods, and named PartsTri and PartsTetra, respectively. Table 1 shows that the level of amino acid-dependent clustering on PartsTetra (Fig. 2A) and PartsTri (data not shown) is slightly higher than on the previous Penta for most amino acids. Figure 2B shows that amino acid-dependent clustering is simpler, and minor territories and satellite spots are less evident than in Fig. 1B and Supplementary S2B, supporting the view that the level of amino acid-dependent self-organization has increased in Parts-BLSOMs (Table 1). Similar results were obtained for PartsTri (Supplementary Fig. S5). Figure 2C presents the 3D view of the number of Arg or Met tDNAs in each lattice point on PartsTri. The merit of Parts-BLSOM is not only the slight increase of amino-acid dependent clustering, but also that it can provide the following strategies for instructive knowledge discovery. (The U-matrix shown in Fig. 2D is explained later.)

**Strategies for revealing functionally and/or structurally important domains** Functionally important sequences, such as identity elements, have been experimentally proven to differ often in sequence location among tRNAs belonging one isoacceptor group and additionally among phylogenetic groups (Hou and Schimmel, 1988; McClain and Foss, 1988; Marck and Grosjean, 2002; Ardell, 2010). For an *in silico* prediction of their locations, we have constructed PartsTri and PartsTetra, in which one of six functional parts is omitted. The per-
centages of pure lattice points (i.e., lattice points containing tDNAs only of one amino acid) on PartsTetra are listed in Fig. 3A. When omitting the anticodon arm, the separation level decreases significantly (to < 75%) for His, Lys and Trp, and less significantly for Arg, Asn, Gln, Ile, Met, Phe, Pro, Thr and Val. This reduction probably reflects the contribution level of each part for amino acid-dependent clustering and is consistent with the locations of identity elements experimentally proven for various model organisms (Normanly and Abelson, 1989; Ibba and Söll, 2000; Marck and Grosjean, 2002; Ardell, 2010).

An alternative analysis is tri- and tetranucleotide BLSOM (Tri and Tetra) constructed separately for each part. Figure 3B shows the percentages of pure lattice points for each amino acid on Tetra for each part. The anticodon arm gives a good separation (> 90%) for His, Lys and Trp, and less significantly for Arg, Asn, Gln, Ile, Met, Phe, Pro, Thr and Val. This reduction probably reflects the contribution level of each part for amino acid-dependent clustering and is consistent with the locations of identity elements experimentally proven for various model organisms (Normanly and Abelson, 1989; Ibba and Söll, 2000; Marck and Grosjean, 2002; Ardell, 2010).

An alternative analysis is tri- and tetranucleotide BLSOM (Tri and Tetra) constructed separately for each part. The anticodon arm gives a good separation (> 90%) for all amino acids (His, Lys and Trp, and less significantly for Arg, Asn, Gln, Ile, Met, Phe, Pro, Thr and Val). This reduction probably reflects the contribution level of each part for amino acid-dependent clustering and is consistent with the locations of identity elements experimentally proven for various model organisms (Normanly and Abelson, 1989; Ibba and Söll, 2000; Marck and Grosjean, 2002; Ardell, 2010).

An alternative analysis is tri- and tetranucleotide BLSOM (Tri and Tetra) constructed separately for each part. The anticodon arm gives a good separation (> 90%) for all amino acids (His, Lys and Trp, and less significantly for Arg, Asn, Gln, Ile, Met, Phe, Pro, Thr and Val). This reduction probably reflects the contribution level of each part for amino acid-dependent clustering and is consistent with the locations of identity elements experimentally proven for various model organisms (Normanly and Abelson, 1989; Ibba and Söll, 2000; Marck and Grosjean, 2002; Ardell, 2010).

A major source for finding tDNAs is the massive number of metagenomic sequences, derived from a wide range of environmental and clinical samples, that have been compiled in international DNA data banks (INSDC: DDBJ/ENA/NCBI). Since metagenomic sequences have attracted broad scientific, industrial and medical interest, tRNADB-CE has included tDNAs obtained from...
metagenomic sequences (abbreviated to metagenomic tDNAs); these metagenomic tDNAs are analyzed here with BLSOM. Since metagenomic sequences are probably derived not only from bacteria but also archaea and fungi, we have constructed PartsTetra with species-unknown metagenomic tDNAs plus species-known bacterial, archaeal and fungal tDNAs, comprising 0.6 million tDNAs in total (Both in Fig. 4A).

Species-unknown metagenomic and species-known microbial tDNAs are visualized separately in Metagenome and Known in Fig. 4A. Amino acid-dependent clustering is apparent, but separation patterns are more complex than those for species-known bacterial tDNAs listed in Fig. 1A, and more black lattice points appear in Fig. 4A than in Fig. 1A. The majority of black lattice points are observed for metagenomic tDNAs (Metagenome in Fig. 4A), with a minority also observed for species-known tDNAs (Known in Fig. 4A). Figure 4B separately marks lattice points containing metagenomic and species-known tDNAs of Ala and Asn (for other amino acids, see Supplementary Fig. S5). Detailed inspection of the species-known tDNAs belonging to black lattice points in Fig. 4A has revealed these to be primarily archaeal and fungal tDNAs, showing that BLSOM has separated archaeal and fungal tDNAs from bacterial tDNAs and that a plenty level of metagenomic tDNAs is derived from archaea and fungi. The observation that a large portion of archaeal and fungal tDNAs are located in black lattice points indicates that their self-organization depends largely on sequence characteristics that distinguish them from bacterial tDNAs, rather than on distinctions between amino acids. To study amino acid-dependent clustering of archaeal and fungal tDNAs, BLSOMs have to be constructed only for archaeal and fungal tDNAs.

The vertical bar in Fig. 4C presents the number of metagenomic and species-known tDNAs of Ala and Asn (for other amino acids, see Supplementary Fig. S5). Detailed inspection of the species-known tDNAs belonging to black lattice points in Fig. 4A has revealed these to be primarily archaeal and fungal tDNAs, showing that BLSOM has separated archaeal and fungal tDNAs from bacterial tDNAs and that a plenty level of metagenomic tDNAs is derived from archaea and fungi. The observation that a large portion of archaeal and fungal tDNAs are located in black lattice points indicates that their self-organization depends largely on sequence characteristics that distinguish them from bacterial tDNAs, rather than on distinctions between amino acids. To study amino acid-dependent clustering of archaeal and fungal tDNAs, BLSOMs have to be constructed only for archaeal and fungal tDNAs.

The vertical bar in Fig. 4C presents the number of metagenomic and species-known tDNAs of Ala and Asn. Figure 4C shows, more clearly than Fig. 4B, that the metagenomic tDNAs located outside the major ter-

---

**Fig. 3.** Contribution level of each functional part to amino acid-dependent clustering. (A) The percentages of pure lattice points for each amino acid on PartsTetra, in which one functional part is omitted, are presented by vertical bars colored as follows: no omission ( ), omission of 5’ acceptor ( ), D-arm ( ), anticodon arm ( ), V-arm ( ), T-arm ( ), and 3’ acceptor ( ). (B) The percentages of pure lattice points on Tetra for all parts and each part are presented for each amino acid by vertical bars colored as follows: 5’ acceptor ( ), D-arm ( ), anticodon arm ( ), V-arm ( ), T-arm ( ), and 3’ acceptor ( ).
ritories and primarily representing archaeal and fungal tDNAs are more abundant than species-known tDNAs. Furthermore, the locations of very high peaks in the major territories differ between metagenomic and species-known tDNAs. This can be more clearly shown by the following analysis of tDNAs of each amino acid.

**BLSOM for each amino acid** Dick et al. (2009) successfully applied the U-matrix method (Ultsch, 1993) of an oligonucleotide-SOM to the phylogenetic clustering of environmental metagenomic sequences. The U-matrix presented in Fig. 2D visualizes the dissimilarity level of oligonucleotide composition between neighboring lattice points as a grayness level; dark gray lines correspond primarily to borders between different amino acid territories, showing a clear dissimilarity of oligonucleotide compositions in tDNAs related to different amino acids. In addition, even within a major territory of one amino acid, many partitions surrounded by pale gray lines are observed and appear to be primarily attributable to phylogenetic differences. To investigate the phylotype-dependent separation in more detail, we constructed a BLSOM for each amino acid for species-known plus species-unknown tDNAs (Fig. 5). On the All panel in Fig. 5A, lattice points containing Leu tDNAs from only bacterial, archaeal, fungal or metagenomic sequences are colored in blue, red, green or gray, respectively; those containing tDNAs from more than one category are marked in black. On the Bacteria, Archaea and Fungi panels, lattice points containing metagenomic tDNAs (gray) plus bacterial, archaeal or fungal tDNAs are separately colored, as described for the All panel. A large portion of lattice points on the Bacteria panel are marked in black, showing that many metagenomic tDNAs are clustered (self-organized) with known bacterial tDNAs, predicting phylogenetic attribution of metagenomic tDNAs. On the Bacteria panel, some clear gray contiguous areas contain metagenomic (but not bacterial) tDNAs, and some gray areas contain archaeal and fungal tDNAs (black on the Archaea or Fungi panel), providing phylogenetic attribution of these metagenomic tDNAs. On the U-matrix panel in Fig. 5A, many white or pale gray areas are surrounded by dark gray circles. White and pale gray on the U-matrix show similar oligonucleotide compositions between neighboring lattice points, i.e., between tDNAs located in neighboring lattice points. Therefore, metagenomic tDNAs within a white and pale gray zone surrounded by a dark gray circle can be phylogenetically assigned by referring to species-known tDNAs colocalizing in this zone, as described by Dick et al. (2009). A significant portion of metagenomic tDNAs is also located apart from species-known tDNAs (gray contiguous areas in the All panel), showing these tDNAs to be derived mainly from poorly studied genomes that exist in novel...
AI approach for tDNA studies

environmental sources.

The vertical bars in Fig. 5B present the number of species-known and metagenomic tDNAs of three amino acids. The locations of even very high peaks differ between species-known and metagenomic tDNAs. We next focus on the very high peaks observed for metagenomic tDNAs. In the case of Ala, the highest peak (marked 1) contains 1,205 metagenomic tDNAs primarily obtained from marine samples plus two tDNAs of Candidatus Pelagibacter (an oceanic carbon-recycling bacterium); peak 2 contains 171 tDNAs primarily obtained from hot spring samples plus two tDNAs of Synechococcus sp. JA-3-3Ab (Cyanobacteria bacterium Yellowstone A-Prime); and peak 3 contains 159 marine metagenomic tDNAs but no species-known tDNAs. In the case of Pro, peak 1 contains 244 marine metagenomic tDNAs plus one Candidatus Pelagibacter tDNA; peak 2 contains 152 marine metagenomic tDNAs plus nine Chlorobi tDNAs; and peak 3 contains 287 marine tDNAs but no species-known tDNAs. In the case of Leu, peak 1 contains 196 marine metagenomic tDNAs plus four Chlorobi tDNAs; peaks 2 and 3 contain 316 and 309 marine metagenomic tDNAs but no species-known tDNAs. Results for each amino acid can provide phylogenetic information about metagenomic sequences and assign novel tDNAs with new sequence characteristics. Importantly, tDNAs that form peaks composed of multiple tDNAs should be reliable tDNAs even though they have noncanonical characteristics. This can specify a large number of new types of tDNAs, which have been poorly characterized. The analysis of big sequence data, such as those obtained from metagenomic samples, can provide this type of novel information.

Fig. 5. PartsTetra for metagenomic plus species-known microbial tDNAs of one amino acid. (A) Two-dimensional view. All: lattice points containing tDNAs derived from only bacterial, archaeal, fungal or metagenomic sequences are colored in blue, red, green or gray, respectively, and those containing tDNAs from more than one category are marked in black. Bacteria, Archaea and Fungi: lattice points containing metagenomic tDNAs plus bacterial, archaeal or fungal tDNAs are separately colored as described for the All panel. U-matrix is presented as described in Fig. 2D. (B) Three-dimensional view. The number of metagenomic and species-known tDNAs in each lattice point on PartsTetra is represented by the height of the vertical bars.
DISCUSSION

Characteristics of unsupervised data mining The present in silico findings obtained from more than 7,000 bacterial genomes, for most of which molecular studies other than DNA sequencing are lacking, can be connected with molecular mechanisms that have been experimentally proven for a limited number of model organisms, such as those reviewed by Marck and Grosjean (2002). In their review, 50 genomes were selected to avoid overrepresentation of organisms that are, phylogenetically, too closely related to each other and to span the widest range of living species; over 4,000 tDNAs were extracted, analyzed and compared. In our study, as is typical for big data analyses, we have used all available bacterial data without particular filtration processes and thus incorporated genomes of closely related bacteria, including those of different strains of one species; in total, 0.4 million tDNAs from more than 7,000 bacterial genomes. These two distinct analyses are complementary to each other, and BLSOM can clarify the range of species to which the experimentally proven mechanisms are applicable and can point out inapplicable phylotypes.

Phylogenetic markers useful for short metagene- nomic sequences When searching for a genome of particular interest (e.g., for industrial usability) by surveying a massive number of short metagenomic sequences, phylogenetic marker tDNAs should be very useful because of their short length. Our group has started to search for tDNAs that are useful as phylogenetic markers, especially for rare genomes, and will publish such markers in tRNADB-CE. The present study shows that the BLSOM with species-known tDNAs plus species-unknown metagenomic tDNAs (Fig. 4) can provide a tool for studying a microbial community in an ecosystem. When analyzing a dataset composed mainly of sequences shorter than 100 bp, this strategy is useful since conventional phylogenetic tree methods cannot be properly applied to most short sequences; it is impossible to construct reliable phylogenetic trees for most of these short sequences. If the data-set is composed mainly of sequences longer than 500 bp, BLSOMs with tri- and tetranucleotide compositions in all genomic fragments should be more suitable than tDNA-BLSOM, because all genomic sequences are informative (Abe et al., 2005; Nakao et al., 2013).

It should also be mentioned that horizontal gene transfers between different species are a general characteristic of microbial genomes. Therefore, we may not find phylogenetic markers with 100% accuracy, because informatics methods, including sequence homology searches, most likely assign the horizontally transferred genes to the donor and not the recipient genome. When noncanonical tDNAs are found concurrently in restricted members of phylogenetically distant groups (e.g., different classes and families), the genes may represent horizontally transferred genes or products of convergent evolution. Use of phylogenetic marker tDNAs must take these points into consideration.

Strategies for enhancing the accuracy of a large-scale tRNA database As described in Materials and Methods, to enhance accuracy in compiling tDNAs in tRNADB-CE, three computer programs have been used in combination, since their algorithms partially differ and render somewhat different results (Abe et al., 2011). For tDNA candidates predicted by only one or two programs, experts in tRNA experimental research manually checked them. Searching for the minimum anticodon set for a completely sequenced genome (Osawa, 1995; Marck and Grosjean, 2002) is an important check process (Abe et al., 2011). Another manual check applicable even to partially sequenced genomes is to examine whether the candidates have been found iteratively in closely related species; this process has become increasingly useful because the genomes of many closely related species and even of different strains belonging to one species have been sequenced. When the same or almost the same noncanonical sequences were found repeatedly, the tDNAs were included in the Reliable tRNA category (Abe et al., 2011), based on the knowledge that functionally important genes have been stably maintained throughout evolution. Furthermore, accumulation of a large number of metagenomic sequences has progressively increased the reliability of this verification strategy, by pointing to promising characteristics of big data.

For creating a large-scale and high-quality database, it is important to find errors that have slipped into the database, including those caused by DNA sequencing errors. As mentioned above, tDNAs found concordantly by all three programs have been stored in tRNADB-CE after brief anticodon checking. While this automatic compilation is indispensable for surveying the huge number of genomic sequences accumulated in INSDC, a new strategy for identifying erroneous cases is required for quality enhancement. Orphan tDNAs located outside correspondent amino acid territories on BLSOM are candidates for erroneous cases, because the present data mining method has pointed out their sequence irregularity. Such cases should be manually checked by experts, even when three computer programs have concordantly assigned them. In contrast, if tiny spots outside their correspondent major territories harbor multiple tDNAs (e.g., peaks in Figs. 1C, 2C, 4C and 5B), especially when the tDNAs are derived from phylogenetically related species, they are likely to represent real tDNAs, even though they have noncanonical sequence characteristics. These tDNAs will become phylogenetic markers with high specificity for their respective phylotypes. In tRNADB-CE, such tDNAs are noted in a column that has been provided.
for comments on each tDNA (Abe et al., 2011).

When constructing tRNADB-CE, we encountered a significant number of cases where all three programs predicted different functional segmentation for one tDNA candidate although the programs concordantly assign it as tDNA. This may indicate a demerit of Parts-BLSOM because it requires information concerning functional segmentation. When analyzing metagenomic sequences, which probably include various novel genomes, the ordinary BLSOM may be useful because it does not require this prior information. Undoubtedly, their combinatorial use should be a better choice, and this will predict the proper functional segmentation of the tDNA.

CONCLUSION

Unsupervised data mining, which can extract a wide range of knowledge from big data without prior knowledge or particular models, is well-timed in the era of big sequence data accumulation in genome research. Importantly, unsupervised data mining such as BLSOM can provide the least expected knowledge. In addition, to gain a wide range of knowledge efficiently from big data, it is important to view all data simultaneously on one map and to focus on a specific data category by using strong visualization power. Oligonucleotide BLSOM, which can analyze more than ten million sequences at once, is suitable for unveiling novel knowledge hidden within big sequence data, providing a timely tool for a wide range of genome research, which has been enabled by the remarkable progress of high-throughput sequencing technology.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

We thank Drs. Hachiro Inokuchi and Yuko Yamada of Nagahama Institute of Bio-Science and Technology and Dr. Akira Muto of Hiroaki University for valuable discussions. We also thank Mr. Yudai Miyake of Nagahama Institute of Bio-Science and Technology for preliminary BLSOM analyses of tDNAs. This work was supported by Grants-in-Aid for Scientific Research (C: nos. 23500371 and 28330334) and a Grant-in-Aid for Publication of Scientific Research Results (no. 15HP7006), from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and by the NIG Collaborative Research Program (A). The computation was done in part with the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

REFERENCES


