DNA Research **18**, 125–136, (2011) doi:10.1093/dnares/dsr005
Advance Access Publication: 28 March 2011

Prediction of Directional Changes of Influenza A Virus Genome Sequences with Emphasis on Pandemic H1N1/09 as a Model Case

Yuki Iwasaki†, Takashi Abe*,†, Kennosuke Wada, Masae Itoh, and Toshimichi Ikemura

Nagahama Institute of Bio-Science and Technology, Tamura-cho 1266, Nagahama-shi, Shiga-ken 526-0829, Japan

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

Edited by Katsumi Isono
(Received 24 December 2010; accepted 27 February 2011)

Abstract

Influenza virus poses a significant threat to public health, as exemplified by the recent introduction of the new pandemic strain H1N1/09 into human populations. Pandemics have been initiated by the occurrence of novel changes in animal sources that eventually adapt to human. One important issue in studies of viral genomes, particularly those of influenza virus, is to predict possible changes in genomic sequence that will become hazardous. We previously established a clustering method termed 'BLSOM' (batch-learning self-organizing map) that does not depend on sequence alignment and can characterize and compare even 1 million genomic sequences in one run. Strategies for comparing a vast number of genomic sequences simultaneously become increasingly important in genome studies because of remarkable progresses in nucleotide sequencing. In this study, we have constructed BLSOMs based on the oligonucleotide and codon composition of all influenza A viral strains available. Without prior information with regard to their hosts, sequences derived from strains isolated from avian or human sources were successfully clustered according to the hosts. Notably, the pandemic H1N1/09 strains have oligonucleotide and codon compositions that are clearly different from those of human seasonal influenza A strains. This enables us to infer future directional changes in the influenza A viral genome.

Key words: influenza virus; pandemic; self-organizing map; oligonucleotide composition; codon usage

1. Introduction

One important issue for bioinformatics studies of virus genomes, especially of influenza viruses, is to develop a strategy to predict genome sequence changes that will become hazardous.[1–3] The phylogenetic analysis based on sequence homology searches is a well-established and an irreplaceably important method for studying genomic sequences.[4,5] However, it inevitably depends on alignments of sequences, which is potentially error-prone and troublesome especially for distantly related sequences. This difficulty becomes increasingly evident as the number of sequences obtained from a wide range of species, including novel species, increases dramatically because of the remarkable progress of the high-throughput DNA sequencing methods. To address the difficulty and complement the sequence homology searches, we previously established an alignment-free clustering method BLSOM (batch-learning self-organizing map)[6,7] that can analyse a million sequences simultaneously on the basis of SOM originally developed by Kohonen and his colleagues.[8,9] SOM is a powerful unsupervised clustering method that provides an efficient interpretation of complex data, using visualization on one map. Unlike the SOM, the BLSOM depends on neither the data-input order nor the initial conditions, and therefore, is suitable for studying genomic sequences.

G + C% has been used for a long period as a fundamental parameter for phylogenetic classification of microbial genomes, including viral genomes, but the G + C% is apparently too simple a parameter to

† These two authors contributed equally to this work.

differentiate a wide variety of known microbial genomes. Oligonucleotide composition, on the other hand, can be used to distinguish the species even with the same G + C%, because the oligonucleotide composition varies significantly among the genomes and is called 'genome signature'.[10] When we constructed a BLSOM for oligonucleotide frequencies in fragment sequences (e.g. 10 kb) from a wide variety of microbial species, sequences were clustered (self-organized) according to species. BLSOMs could recognize and visualize species-specific characteristics of oligonucleotide composition.[6,7] The alignment-free clustering methods, BLSOM[11–13] and SOM,[14,15] have been successfully applied to phylogenetic classification of a large number of microbial sequences obtained by metagenome studies of environmental and clinical samples. Here, we have analysed influenza A viruses, including those of the pandemic H1N1/09,[16–18] with BLSOM and developed a strategy for predicting directional sequence changes in influenza A viruses. This strategy should be widely applicable for other zoonotic viruses. Introduction of BLSOM into viral genome studies can undoubtedly provide a powerful tool for efficiently extracting profound knowledge from a vast number of viral genomic sequences obtained by high-throughput sequencers currently available.

2. Materials and methods

2.1. Viral genome sequences

A total of 43 831 virus sequences analysed in Fig. 1A were obtained from the DDBJ Genome Information Broker for Viruses (GIB-V; http://gib-v.genes.nig.ac.jp/),[19] and a total of 59 512 segment sequences derived from 7439 influenza A virus strains were obtained from the NCBI Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/).[20]

2.2. Batch-learning self-organizing map

SOM is an unsupervised neural network algorithm that implements a characteristic non-linear projection from the high-dimensional space of input data onto a two-dimensional array of weight vectors.[8,9] We modified the conventional SOM for genome informatics to make the learning process and resulting map independent of the order of data input, and established a BLSOM.[6,7,21] The initial weight vectors were defined by principal component analysis instead of random values. BLSOM learning for oligonucleotide composition was conducted as described previously.[6,7] BLSOM learning for synonymous codon usage and visualization of diagnostic codons for the category separation were conducted as described by Kanaya *et al*.[21] BLSOM program was obtained from UNTROD, Inc. (y_wada@nagahama-i-bio.ac.jp).

3. Results and discussion

3.1. Oligonucleotide BLSOM for all virus genome sequences

To test the clustering powers of BLSOM for large numbers of sequences from a wide variety of virus genomes, we initially constructed BLSOM with tri- and tetranucleotide frequencies in all 1-kb fragment sequences (ca. 200 000 sequences) from 42 957 virus genomes, which have been compiled and classified into 79 phylogenetic families by DDBJ GIB-V[19] (Fig. 1A). Lattice points that contained sequences from one phylogenetic family are indicated in colour, and those that included sequences from more than one family are indicated in black. A major portion was coloured, showing a major portion of sequences to be self-organized according to phylotype; ~80 and 86% of lattice points had sequences from one phylotype on tri- and tetra-BLSOMs, respectively. Notably, no information in regard to phylotype was given during the BLSOM calculation. Tetra-BLSOM for 0.5-kb sequences (ca. 4 000 000 sequences) also showed a clear clustering according to phylotype, with a slight reduction (~5%) in the separation (Fig. 1A); tri-BLSOM for 0.5-kb sequences is presented in Supplementary Fig. S1.

3.2. Oligonucleotide BLSOM for influenza A virus genomes

We next analysed influenza A virus sequences available from Influenza Virus Resource[20] in NCBI. Influenza A virus genome is composed of eight segments, each of which encodes primarily one protein. Genome sequences from a total of approximately 7400 strains were available even when we focused on strains for which all eight segments were sequenced, and approximately 2300 strains corresponded to the new pandemic H1N1/09. The direct target of natural selection is a virion containing a full set of the eight genome segments, and analyses at the genome level should provide valuable information for characterizing individual strains. Viruses are inevitably dependent on many host factors for their growth (e.g. pools of nucleotides, amino acids and tRNAs), and at the same time have to escape from antiviral host mechanisms such as antibodies, cytotoxic T cells, interferons, and RNA interferences.[22–25] Actually, mononucleotide compositions and codon usage biases in influenza A virus genomes were shown to differ between strains isolated from human and avian sources.[26,27]

To clarify the host-dependent characteristics of influenza A virus sequences with BLSOM, di-, tri-, or tetranucleotide frequencies in eight genome segments of virus strains were summed up for each

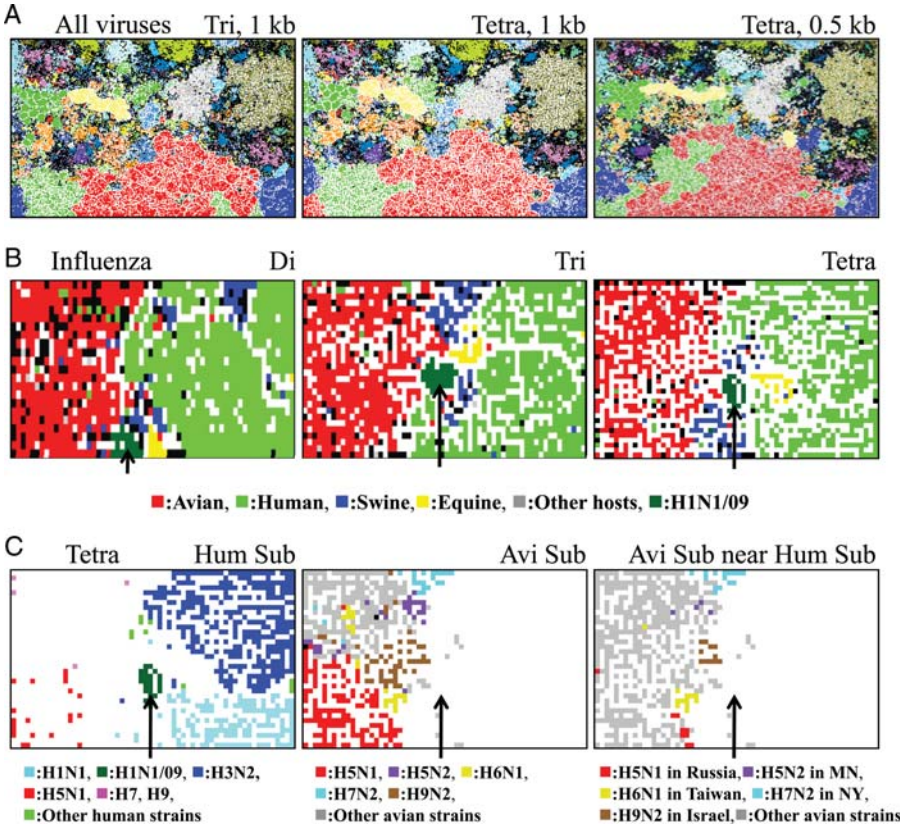

Figure 1. Oligonucleotide-BLSOMs for virus genome sequences. (A) BLSOM was constructed for tri- or tetranucleotide composition (Tri or Tetra) in 1- or 0.5-kb fragment sequences from 42 957 virus genomes. Lattice points that include sequences from more than one phylogenetic family were indicated in black, and those containing sequences from a single phylotype were indicated in a colour representing the phylotype described by the DDBJ GIB-V (see the legend of Supplementary Fig. S1 for the phylotypes). (B) Influenza Di, Tri, or Tetra: di-, tri-, or tetranucleotide BLSOM was constructed for 5350 strains of influenza A virus including 167 H1N1/09 strains isolated in the early stage (April and May) of the 2009 pandemic. Lattice points containing sequences from strains isolated from more than one host were indicated in black, and those containing sequences from one host were indicated in a colour representing the host shown at the bottom of (B): avian (1948 strains), human (2788 strains), swine (249 strains), equine (68 strains), and other hosts (130 strains); human H1N1/09 (167 strains) was separately coloured and indicated with an arrow. (C) Tetra Hum Sub (human subtype): on the tetra-BLSOM presented in (B), each human virus subtype was specified in a colour representing the subtype (H1N1, H1N1/09, H2N2, H3N2, H5N1, H7, H9) shown at the bottom of Tetra Hum Sub, and territories of other hosts were achromatic. Avi Sub (avian subtype): each avian subtype was specified in a colour representing the subtype (H5N1, H5N2, H6N1, H7N2, H9N2, others) shown at the bottom of Avi Sub, and territories of other hosts were achromatic. The zone representing the minor human territory representing H1N1/09 (achromatic) was indicated with an arrow to help recognize its position for reference. Avi Sub near Hum Sub : avian subtype strains that were in a close proximity to human, swine, or pandemic H1N1/09 territory were specified by a colour representing both the subtype and the geographical information of isolation (H5N1 isolated in Russia, H5N2 in Minnesota (MN), H6N1 in Taiwan, H7N2 in New York (NY), H9N2 in Israel, other avian strains) shown at the bottom of Avi Sub near Hum Sub.

strain, and BLSOM was constructed with the summed frequency for each strain. The merit of the present sequence data set of viral genomes was to include those from the H1N1/09 strains isolated in an early stage of the new pandemic, which was caused by a new virus invasion from other hosts. To specifically study sequence characteristics of the strains isolated in the early stages, we first analysed the data set only including approximately 170 H1N1/09 strains that were isolated from April to May in 2009 (Fig. 1B). Lattice points containing virus strains isolated from one host species were indicated in a colour representing the host and those containing strains isolated from more than one host were in black. Without information with regard to the host during the BLSOM calculation, strains isolated from avian (red) or human (light green) were clustered, forming a large contiguous territory on each BLSOM. This showed host-specific characteristics of oligonucleotide composition and was consistent with the observation of Rabadan *et al.*[26] that mononucleotide compositions in influenza A virus genes differed between strains isolated from human and avian sources.

Notably, a minor human territory (dark green and indicated by an arrow) appeared, which was separated from the major human territory representing human seasonal strains and surrounded by avian,

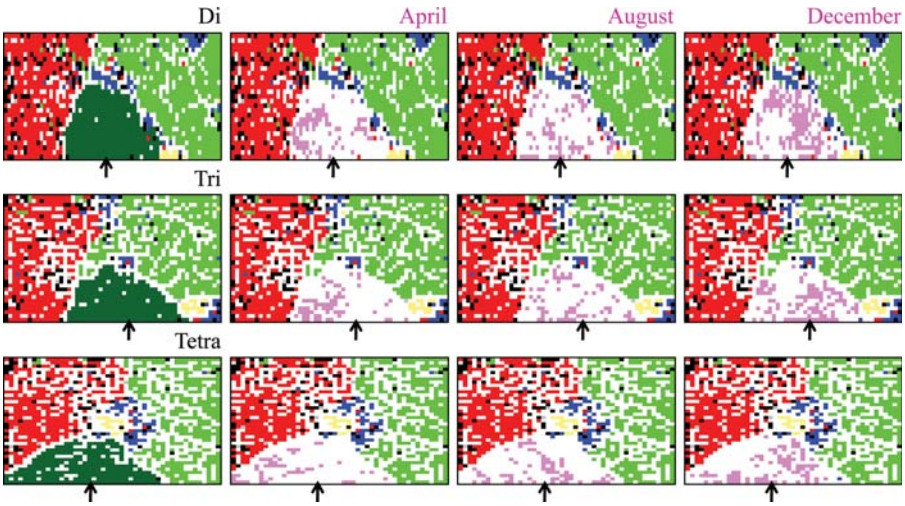

Figure 2. Oligonucleotide-BLSOMs for all influenza A virus sequences. Di, Tri, or Tetra: BLSOMs were constructed for di-, tri-, or tetranucleotide composition in genome sequences from influenza A viruses including all 2256 H1N1/09 strains (arrowed), and lattice points were coloured as described in Fig. 1B. April, August, or December: lattice points containing the H1N1/09 strains isolated in April, August, or December were coloured in pink in the H1N1/09 territory, and lattice points in the territories other than the H1N1/09 territory were coloured as described in Fig. 1B.

equine, and swine territories on each BLSOM. This minor human territory was composed of the pandemic H1N1/09 strains, which have resulted from genetic re-assortment between the recently circulating swine H1 viruses in North America and the avian-like swine viruses in Europe.[17,18]

To further investigate human virus subtypes other than the H1N1/09, lattice points that contained human seasonal viruses of one subtype on the tetra-BLSOM (Tetra in Fig. 1B) were specified with one colour representing the subtype (Hum Sub in Fig. 1C). Human seasonal H1N1 (light blue) and H3N2 (dark blue) were clearly separated from each other. In contrast to the compact minor territory of H1N1/09 (dark green, arrowed), human H5N1 strains (red) were rather scattered within the avian territory (achromatic in Hum Sub in Fig. 1C). This should show that these human H5N1 strains jumped to humans but were not able to spread from human to human,[24] and therefore, they had characteristics of avian viruses. These human H5N1 strains were more separated from the swine and human territories than H1N1/09 strains, and this difference may relate with their infection power in the human and swine populations.

In order to identify avian strains that were located near the swine and human territories, we next marked lattice points containing sequences from one avian subtype by one colour (Avi Sub in Fig. 1C). Avian subtypes were clearly separated from each other. Some avian strains were in closer proximity to human or swine territory than the human H5N1 strains already known (Avi Sub near Hum Sub in Fig. 1C); for example, H5N1 isolated in Russia (red), H6N1 in Taiwan (yellow), H9N2 in Israel (brown), H5N2 in Minnesota (violet), and H7N2 in New York (light blue). These strains should have oligonucleotide compositions more similar to those of human and/or swine viruses than the known human H5N1 strains. This similarity may reflect in part their evolutionary histories, during which the virus themselves or their segments have changed hosts. If so, the similarity is of particular interest with regard to potential infection powers in human and/or swine populations when these strains invade these populations. The finding that the H1N1/09 strains were located in a closer vicinity of both swine and human territories than the H5N1 strains isolated from humans on BLSOM was consistent with this view, although much more data of new pandemics of various subtypes should be needed before proving this view. Even when analysing a large number of strains, BLSOM could effectively visualize and characterize, from various angles, the strains with peculiar characteristics, on which experimental and/or medical research groups will focus their interests. In the present study, strains isolated from various avian hosts (e.g. chicken, duck, mallard and turkey) were grouped into one category. In the near future, a large number of strains isolated from various avian hosts will be sequenced because of importance both in medicine and in chicken and cattle industries, and difference in oligonucleotide composition might be detectable for different avian hosts on BLSOM.

In order to clarify an overall feature of the H1N1/09 pandemic, we next investigated the data set including all 2300 H1N1/09 strains available (Fig. 2). Again, host-dependent separation of strains was observed,

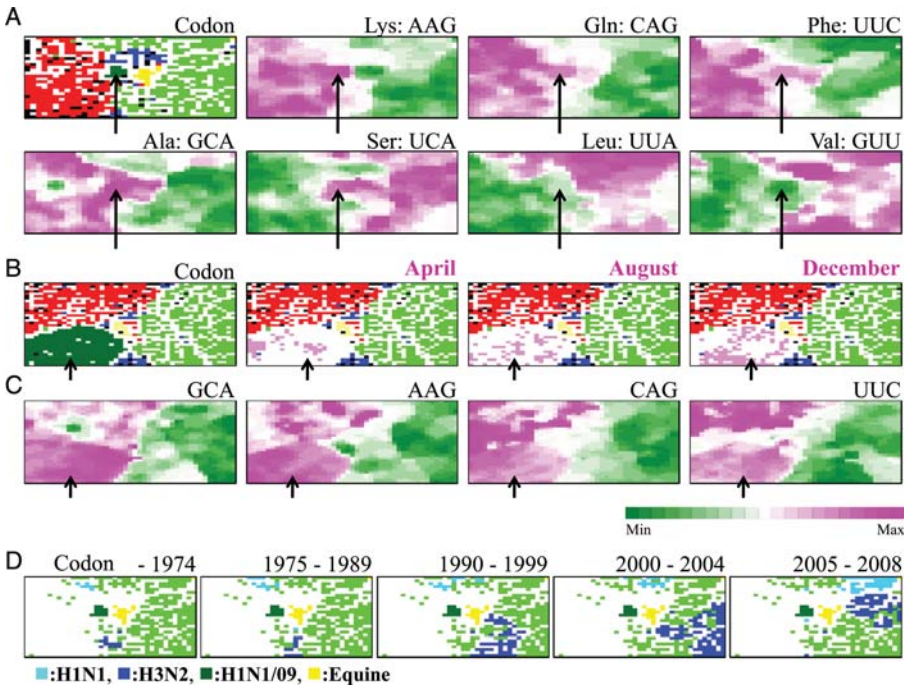

Figure 3. Codon-BLSOM. (A) Codon: BLSOM was constructed for synonymous codon usage in approximately 34 000 genes from influenza A strains including H1N1/09 isolated in the early stage (April and May), as described by Kanaya *et al.*,[21] and lattice points were indicated in a colour representing the host as described in Fig. 1B. Occurrence levels of seven codons, which were diagnostic for host-specific separation, were indicated with different levels of two colours: pink (high), green (low), and achromatic (intermediate) as described by Kanaya *et al.*[21] Because the synonymous codon usage was analysed, two codons belonging to each two-codon box gave the complementary patterns, and therefore, the pattern of one codon was listed for those belonging to two-codon boxes. (B) Codon: BLSOM was constructed for synonymous codon usage in approximately 60 000 genes from influenza A strains including all H1N1/09strains, and lattice points were indicated in a colour representing the host as described in Fig. 1B. (C) Occurrence levels of four diagnostic codons in the Codon-BLSOM presented in (B) were indicated as described in (A). Scale for the occurrence levels in (A) and (C) was indicated in the colours shown at the bottom of (C). (D) Retrospective time-series changes for human subtype strains (H1N1 and H3N2) on the Codon-BLSOM presented in (A). A zone for H1N1/09 or equine strains was additionally marked for reference to help recognize the position in Codon-BLSOM in Fig. 1A. Each strain was indicated in the colours shown at the bottom of (D).

and 2300 H1N1/09 strains formed a large contiguous territory on all BLSOMs (Fig. 2). This showed that even at the late stage of the 2009 pandemic, H1N1/09 strains had sequence characteristics distinct from human seasonal strains. To analyse possible changes that might occur during this pandemic period, strains isolated in the very early stage (April), the middle stage (August), and the later stage (December) in the 2009 pandemic were separately marked in pink in Fig. 2. In the very early stage (April), a major portion of the strains were located in the vicinity of avian and swine territories, but in the later stage (December), strains near the avian territory became a minority and a major portion moved towards the human territory, indicating directional sequence changes. Dispersion within the H1N1/09 territory for strains isolated at the same pandemic stage appeared to reflect primarily their geographical differences.

3.3. *Host-specific codon biases*

Synonymous codon choice sensitively reflects constraints imposed on genome sequences and thus provides a sensitive probe for searching for molecular mechanisms responsible for the constraints, e.g. genome G + C% and tRNA composition in the cases of micro-organisms.[28–31] We previously found that BLSOM efficiently detected species-specific codon-choice patterns of micro-organisms, resulting in self-organization of genes according to microbial species.[21] Furthermore, in the case of genes horizontally transferred relatively recently, codon choice reflected primarily that of the donor, but not the recipient, genome. We next constructed BLSOM for synonymous codon usage in influenza A virus genes including genes from the H1N1/09 strains isolated in the early stage of the 2009 pandemic (Fig. 3A); in order to know codon biases for each strain, codon usages in eight genes were summed for each strain. Human and avian territories were again clearly separated from each other, and human H1N1/09 strains (arrowed in Fig. 3) were again separated from the major human territory and surrounded by avian, equine, and swine territories. Synonymous codon-choice patterns of newly invading viruses, such as

H1N1/09, should be close to those of the original host viruses, at least for a period immediately after the invasion.

Because viruses depend on many cellular factors for their growth, codon choice will most likely shift towards the pattern of seasonal human viruses during many infection cycles among humans. If so, the direction of sequence changes in H1N1/09 in the near future is predictable, and this should be testable because of the high mutation rate and short generation time of influenza A viruses.[24,32] By analysing retrospective changes of codon bias of human seasonal H1 and H3 strains isolated before 2006, Wong *et al*.[27] found the overall reduction of G + C% during course of their evolutions. To examine whether the codon choice in H1N1/09 strains also changed even within the 2009 pandemic, we constructed codon-BLSOM in which genes from all 2300 H1N1/09 strains were included and marked strains isolated in the very early stage (April), the middle stage (August), and the later stage (December) separately in pink (Fig. 3B). A major portion of the strains isolated in the very early stage (April) was located in the vicinity of the avian territory but a major portion in the later stage (December) was apart from the avian territory, as found in the tetra-BLSOM (Fig. 2), supporting the directional sequence changes.

3.4. *Diagnostic codons for host-specific clustering*

We next attempted to identify codons that will change their occurrence levels during the course of the H1N1/09 evolution. BLSOM provides a powerful ability for visualizing diagnostic codons or oligonucleotides that contribute to self-organization of sequences according to hosts.[6,7,21] In Fig. 3A, the frequency of each codon in the representative vector at each lattice point was calculated and sorted according to the frequency, and this rank order was represented at different levels in colours pink (high) and green (low) for each codon, as described previously.[21] Transitions between the high and low ranks often coincided with host territory borders, and seven examples of the codons diagnostic for host separation were presented (Fig. 3A). Clear difference was observed between human and avian territories for these seven codons. Minor differences in the pattern between codons were observed mainly for swine and equine territories. For example, AAG was preferred in the avian territory but neither in a major portion of the human territory nor in the equine territory, which was specified in yellow in the 'Codon' panel in Fig. 3A. In the cases of CAG and UUC, these were preferred in both the avian and equine territories, but not in a major portion of the human territory.

Table 1. Preferred codons and oligonucleotides in avian or human viruses

	Preferred in avian viruses	Preferred in human viruses
Codon	***AAG***, ***CAG***, ***CUC***, CUG, ***GCA***, GCG, GUG, ***UCG***, ***UUC***	***AAA***, ***ACU***, AGA, ***CAA***, ***CCU***, ***GUU***, UCA, ***UUA***, ***UUG***, ***UUU***
Di	***AG***, ***CG***, CU, ***GA***, ***GG***	AA, ***UU***
Tri	ACG, AGG, CAG, ***CCA***, CGU, GAG, GCA, ***GCG***, GGA, ***GUG***, UCC, UCU	AAA, ***AUU***, UAA, UCA, ***UUA***, ***UUU***
Tetra	ACGC, ***ACGG***, ***AGAG***, ***AGCG***, ***CCAC***, CGAG, CGGA, ***CGGC***, CUUC, ***GACU***, GAGC, GAGG, GCAG, GGAG, ***UCUU***, UGUG, ***UUCG***	***AAAA***, AAAU, AAGU, AUUA, AUUU, ***CAAA***, CCAU, CUUU, GGCC, ***GGGG***, ***UGUA***, ***UGUU***, ***UUAA***, UUAU, UUCA, ***UUGU***, ***UUUC***, ***UUUG***, ***UUUU***

To specify the characteristic codon/oligonucleotide preference in the H1N1/09 strains, codons and oligonucleotides preferred in both H1N1/09 and avian strains are indicated in bold italic letters in the column 'Preferred in avian viruses'. Similarly, those preferred in seasonal human viruses but not in H1N1/09 are indicated in bold italic letters in the column 'Preferred in human viruses'.

In Table 1, all diagnostic codons for the separation between human and avian territories were summarized. When we focused on diagnostic codons (Codon in Table 1), one simple tendency was observed. Codons ending with G or C were more favourable for the avian strains than the human strains. This supported the previous observation of codon biases found by Wong *et al*.[27] in their analyses in which individual genes were analysed separately with Corresponding Analysis. Confirmation of the previous finding concerning codon usage with the BLSOM method showed the reliability of this new method. The G + C% effect was most apparent in two-codon boxes (Table 1, Fig. 3A). This was also observed for many codons in four- or six-codon boxes, but there were few exceptional cases, such as codons 'GCA' preferred in the avian territory and 'UUG' preferred in the human territory (Table 1), indicating the presence of constraints other than the G + C% effect.

Notably, for many diagnostic codons, H1N1/09 strains (arrowed) had the avian-type preference rather than the human-type preference (Fig. 3A). In Table 1, to specify this characteristic codon preference in the H1N1/09 strains, codons preferred in both H1N1/09 and avian strains are indicated in bold italic letters in the column 'Preferred in avian viruses', and codons preferred in seasonal human viruses, but not in H1N1/09, are indicated in bold italic letters in the column 'Preferred in human viruses'. Adaptation of codon choice to a new cellular environment (e.g. host body temperature and cellular

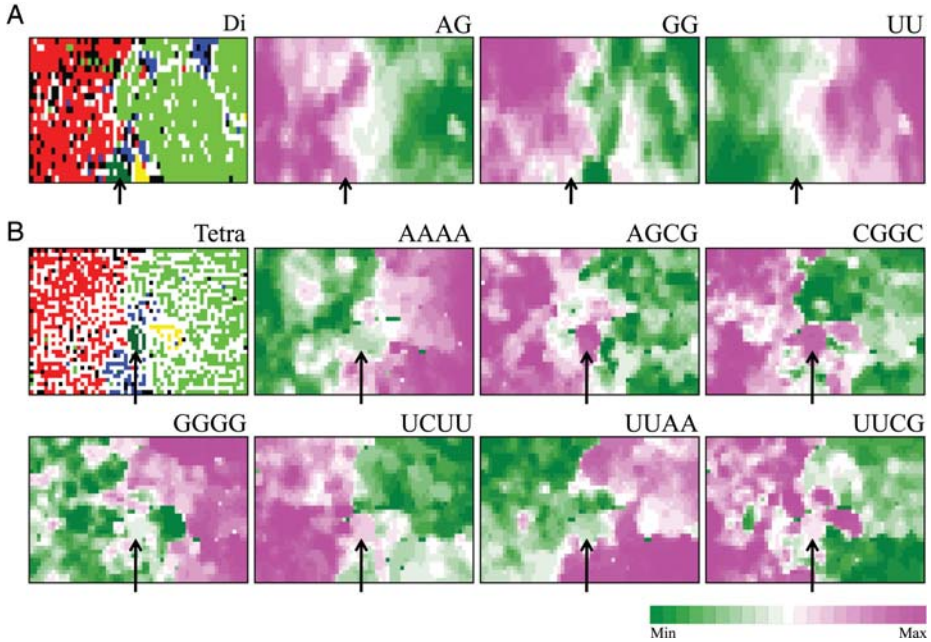

Figure 4. Diagnostic di- or tetranucleotides for host-specific separation. (A) Di: dinucleotide-BLSOM listed in Fig. 1B. Occurrence levels of three dinucleotides, which were diagnostics for host-specific separation, were indicated with different levels of two colours as described in Fig. 3A. (B) Tetra: tetranucleotide-BLSOM listed in Fig. 1B. Occurrence levels of seven diagnostic tetranucleotides were indicated with different levels of two colours as described in Fig. 3A. Scale for the occurrence levels was indicated in the colours shown at the bottom of this figure.

factors) may be a process for a newly invading virus to establish continuous and efficient infection cycles among humans by increasing fitness.

Next, we analysed diagnostic codons for the host-specific separation found in the codon-BLSOM in Fig. 3B, which included all 2300 H1N1/09 strains, and presented four diagnostic examples in Fig. 3C. Codon preference in the H1N1/09 territory (dark green, arrowed) was again similar to that of the avian territory but not of the human territory, showing that even in the late stage of the 2009 pandemic, the codon bias in the H1N1/09 genes did not reach the level found for a major portion of human seasonal virus genes. Uneven pink and green levels were observed within the H1N1/09 territory, and this presumably represented in part the sequence changes accumulated during the 2009 pandemic.

3.5. *Diagnostic oligonucleotides for host-specific clustering*

To identify the oligonucleotides that may change their occurrence levels after the present new virus invasion into human populations, we next identified diagnostic oligonucleotides for host-specific separation on the di-, tri-, and tetranucleotide BLSOMs, which were constructed for the data set including only the 170 H1N1/09 strains isolated in the early stage. Three examples of diagnostic dinucleotides for the host-specific separation on the dinucleotide BLSOM (Di in Fig. 1B) are presented in Fig. 4A, seven examples of diagnostic tetranucleotides are presented in Fig. 4B, and seven examples of diagnostic trinucleotides are presented in Supplementary Fig. S2. All diagnostic di-, tri-, and tetranucleotides for the separation between human and avian territories are summarized in Table 1. Two sorts of distinct tendency were immediately apparent: (i) G- and C-rich oligonucleotides were more favourable in avians than in humans. This G + C% effect was previously reported by Rabadan *et al.*[26] (ii) Oligonucleotides containing AG, CG, or GA dinucleotides were more favourable in avians than in humans. The observation about the CG dinucleotide was consistent with the previous finding by Greenbaum *et al.*[33] that human viruses exhibit a lower CG dinucleotide content than avian viruses. Most of diagnostic di- and trinucleotides could be explained by the above two rules, but there were various exceptional cases for tetranucleotides, indicating the presence of factors other than the two rules. For example, as observed in Fig. 4B, GGGG, a tetranucleotide composed only of G, was preferred mainly in the human territory, whereas UCUU, a tetranucleotide rich in U, was preferred mainly in the avian territory. In the case of interaction of viral components with host factors (e.g. host proteins), oligonucleotide compositions, rather than the mononucleotide composition, will become important, because interactions with host factors primarily depend on oligonucleotide sequences. This should

also be true in considering escape processes from host antiviral mechanisms. BLSOM for oligonucleotide composition should provide valuable information to experimental studies of adaptation mechanisms, such as interactions between viral and host factors.

Notably again, H1N1/09 strains (arrowed in Fig. 4) have characteristics of avian, rather than of human, strains. In the 'Preferred in avian viruses' column in Table 1, oligonucleotides that were preferred in both avian and H1N1/09 strains, but not in human seasonal strains and thus have the potentiality for decrease in H1N1/09 genomes in the near future, are indicated in bold italic letters. Similarly, oligonucleotides that were preferred in human seasonal viruses, but not in H1N1/09 and thus have the potentiality for increase in H1N1/09 genomes in the near future, are indicated in bold italic letters in the 'Preferred in human viruses' column. Searches for nucleotide positions in H1N1/09 genomes with the elevated potentiality of directional changes within antigenic sites,[1–3] antiviral drug-binding sites[34,35] and sites affecting virus growth and virulence[24,32] will provide valuable information for predicting possible H1N1/09 descendant sequences that may present potential hazards. Although protein sequences evolve undoubtedly under the selection on protein sequence, consideration at the nucleotide sequence level is also important. For example, when a certain nucleotide position with an elevated potentiality of the directional changes is the position responsible for coding the amino acid in a certain antigenic site, the probability of this amino acid change is expected to be higher than that for other amino acids in this or other antigenic sites. Additionally, if a certain H1N1/09 strain isolated in a certain geographical area has, in an antiviral drug-binding site, an oligonucleotide with the elevated potentiality of the directional changes, the strains and the area may be counted as a potentially hazardous strain and area. This type of information obtained by informatics methods will become increasingly important in accord with the increase of nucleotide sequences obtained from a wide variety of virus strains.

3.6. *Retrospective time-series changes visualized for human viruses*

Invader viruses will change their sequences on balance between a stochastic process of mutation and selection pressure derived from various constraints, including constraints from hosts. As a model case to study the virus evolution after changing hosts, the 2009 pandemic was very informative because genomes of a large number of strains isolated in various stages, including those isolated in the very early stage, were sequenced. The finding that even in the late stage in the 2009 pandemic the H1N1/09 strains had the characteristics appreciably distinct from human seasonal viruses (Figs. 2 and 3B) indicated that the sequence changes had not reached the presumable equilibrium state, which represented the sequence characteristics common among human seasonal strains. It will take a more extended period to reach the hypothesized equilibrium. Actually, by analysing retrospective changes of codon biases of human seasonal H1 and H3 strains isolated from 1918 to 2006, Wong *et al.*[27] found the overall reduction of G + C% during the course of their evolutions with Corresponding Analysis. We thus examined whether BLSOM could detect the retrospective long-period changes in codon biases in human seasonal H1 and H3 strains. On the codon-BLSOM listed in Fig. 3A, human H1N1 or H3N2 strains that were isolated in the five different periods were separately coloured in light or dark blue for H1N1 or H3N2 strains, respectively (Fig. 3D). Strains isolated before 1975 (shown in the panel '−1974') were located around the border between the human and avian territories, and pandemic descendants (shown in the panels specifying the time period) moved apart from the avian territory (achromatic in Fig. 3D). If human viruses had changed their sequences solely by stochastic processes, pandemic descendants should move primarily in a non-directional way from year to year. Absence of such non-directional movements indicated a directional pressure during the course of establishment of human seasonal strains after the onset of their new pandemics. This finding also showed that BLSOM had a power to visualize evolutionary histories of various subtype viruses, and more importantly to visualize any categories of strains, in which experimental and medical groups will be interested.

3.7. *Segments separately analysed*

At the onset of a new pandemic, re-assortment of virus genome segments in a certain host (e.g. swine) and successive invasion of the new re-assortant into the human population were often essential.[17,18,24,32] Therefore, we next analysed sequences of eight segments separately, which were derived from approximately 5300 strains including 170 H1N1/09 strains isolated in the early stage. The length of the shortest segment (segment 8) is approximately 0.8 kb, and therefore, enough clustering power can be expected as already shown in Fig. 1A. Tetra-BLSOMs for eight segments are presented separately in Fig. 5A; clear clustering of sequences according to host was observed for all segments, and this was true also for di-, tri-, and codon-BLSOMs (data not shown). Segment 2 of H1N1/09 was in close proximity to the human territory, but some other segments

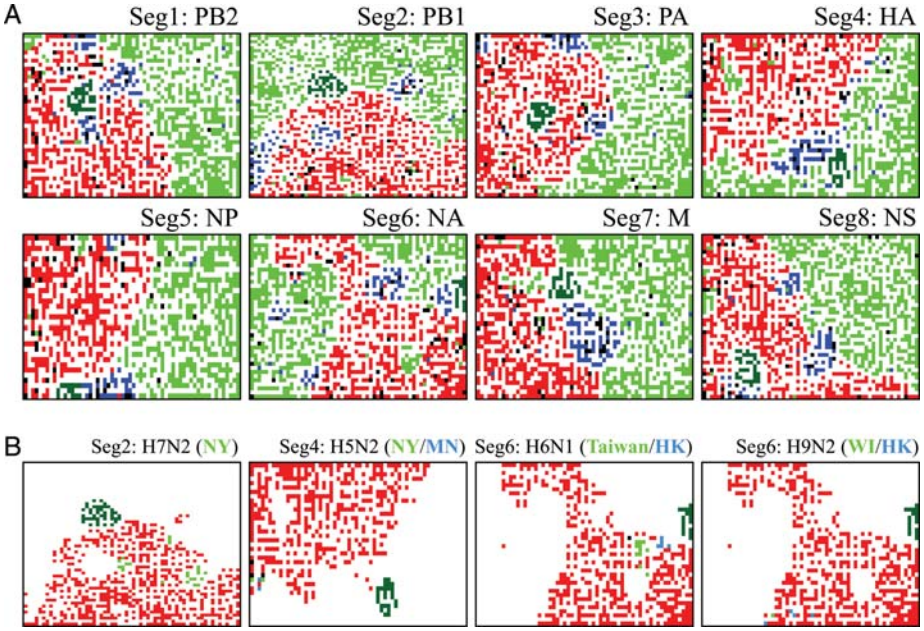

Figure 5. Tetra-BLSOMs for eight genome segments derived from human, avian, and swine viruses. (A) Gene product name was listed along with the segment number. Lattice points were indicated in colours representing the hosts, as described in Fig. 1B, and thus a zone for H1N1/09 was in dark green. (B) Examples of avian strains whose segments were located in close proximity to human and/or swine segments were marked in light green or blue to specify the geographical areas where the subtype strains were isolated: segment 2 of H7N2 strains and segment 4 of H5N2 strains that were isolated in New York (NY), segment 4 of H5N2 strains isolated in Minnesota (MN), H6N1 for segment 6 isolated in Taiwan and Hong Kong (HK), and segment 6 of H9N2 strains isolated in Wisconsin (WI) and Hong Kong (HK). Segments of other avian strains were in red and segments from other hosts were achromatic, but a zone for H1N1/09 was marked in dark green for reference to help recognize the position in (A).

(e.g. segments 1 and 3) were within the avian territory. This similarity of the oligonucleotide composition of H1N1/09 with that of human, swine, or avian viruses was consistent with the findings obtained by conventional phylogenetic studies.[17,18] For example, segment 1 in H1N1/09 was shown to be derived originally from avian strains in the phylogenetic studies, and sequences of segment 1 in the H1N1/09 strains isolated in the early stage were located within the avian territory on BLSOM (Fig. 5A), supporting their evolutionary origin disclosed previously by the phylogenetic studies. During the course of the H1N1/09 evolution, sequence characteristics in segment 1 may have changed towards the characteristics of seasonal human viruses in order to adapt to the new host cellular environments, and this will be examined in the last part of this paper. Sequences of segment 2, which was evolutionarily derived from human strains, were located between human and avian territories, again supporting their evolutionary origin.

Importantly, approximately 5300 sequences were effectively characterized and visualized on one map, supporting efficient knowledge discovery such as those described in what follows. In Fig. 1C (Avi Sub near Hum Sub), we present examples of avian subtype strains that were in close proximity to human and/or swine territories for four subtypes, along with the geographical information of places where the subtype strains were isolated. This provided the information of their similarity at a strain level, potentially unveiling their evolutionary histories at a strain level. In contrast, BLSOM analyses of individual segments presented in Fig. 5A may provide evolutionary histories at a segment level. Visualization and identification of avian- or swine-virus segments whose oligonucleotide and codon compositions were closely related to those of humans should be valuable for predicting candidate strains that may cause pandemics among humans after segment re-assortment. For example, by summarizing potentially hazardous segments, we may specify avian strains that will come to resemble the human or swine strains with re-assortment of only a few segments. This type of information should be very valuable for gaining new perspectives on systematic surveillance of viruses presenting potential hazards. By analysing a large number of virus sequences collectively, BLSOM can efficiently discern such hazardous segments and strains on the basis of its strong visualization power.

Identification of codons or oligonucleotides that are preferred in individual hosts (Table 1) may also provide novel information to aid the design of vaccine candidate strains with high growth rates for supporting high-yield vaccine production and/or the

Table 2. Increase or decrease of codons or oligonucleotides favourable in avian and H1N1/09 viruses

	Segment 1 (Inc/Dec)	Segment 2 (Inc/Dec)	Segment 3 (Inc/Dec)	Segment 4 (Inc/Dec)	Segment 5 (Inc/Dec)	Segment 6 (Inc/Dec)	Segment 7 (Inc/Dec)	Segment 8 (Inc/Dec)
Codon	16/31	27/24	20/19	13/32	9/18	10/12	5/10	6/6
Di	53/95	81/71	61/80	60/68	39/36	45/35	16/22	14/18
Tri	13/26	17/17	23/24	20/19	14/19	14/25	4/10	4/5
Tetra	15/33	22/27	24/34	22/18	18/19	12/12	5/8	12/13
Sum	97/185	147/139	128/157	115/137	80/92	81/84	30/50	36/42

Inc/Dec: the number of increase or decrease of codons or oligonucleotides of attention.

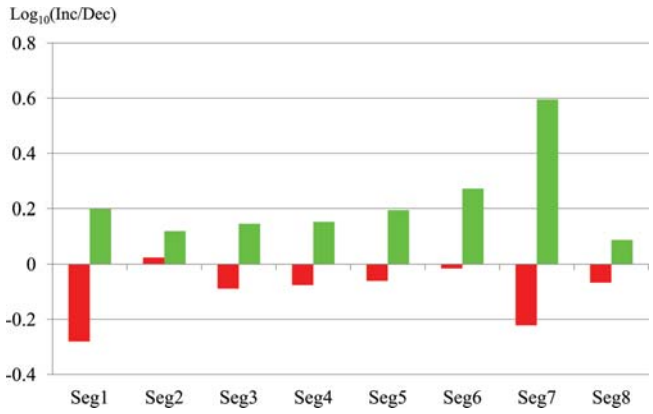

Figure 6. Sequence changes observed in the strains isolated in the latest stage in the 2009 pandemic from an H1N1/09 strain isolated in a very early stage. The ratios 'Inc/Dec' for Sum (summation of di-, tri-, and tetranucleotides and codons) in Tables 2 and 3 were shown in red and green, respectively.

design of generation of recombinant influenza viruses containing various mutations in their genes,[36] which will stably be maintained through infection cycles by avoiding unfavourable codons and oligonucleotides.

3.8. *Confirmation of sequence changes on a basis of sequence homology searches*

To further test the feasibility of the present strategy for predicting directional changes in H1N1/09 sequences, we compared the gene sequences of the strain that was isolated in the very early stage of the pandemic (California/04/2009) with the sequences from approximately 100 H1N1/09 strains that were isolated in the latest stage of the 2009 pandemic (after 1 December 2009). As observed in time-series changes on oligonucleotide- and codon-BLSOMs in Figs. 2 and 3B, a significant level of sequence changes appeared to have accumulated in the H1N1/09 strains during the course of the 2009 pandemic. Next, we analysed base changes at a gene sequence level after alignment with BLAST search, by focusing only on protein-coding sequences because one purpose of the analysis was to recognize changes in codon usage. Furthermore, the exclusion of UTR sequences enabled us to avoid mistakes and/or uncertainties with regard to sequence alignments, which increased significantly for UTR sequences. For each gene of the 100 strains isolated at the latest stage in the 2009 pandemic, nucleotide positions where the base was changed from the sequence of the very early isolated strain were searched for. Then, by keeping focus on the base-changed positions, the number of the codons or oligonucleotides indicated in bold italic letters in the 'Preferred in avian viruses' column in Table 1 (i.e. codons or oligonucleotides predicted to decrees in the near future) that were actually lost or gained in each strain isolated in the latest stage was calculated and summed for the 100 strains (Dec or Inc in Table 2). In this summation, the same mutation found for more than one strain was treated as one change, because it was uncertain whether this base change occurred independently. The number of the lost codons or oligonucleotides of attention exceeded generically the number of those gained (Table 2) (red vertical bars in Fig. 6), supporting the view that the unfavourable codons or oligonucleotides predicted with BLSOM analyses actually had a higher tendency to be lost at the gene

Table 3. Increase or decrease of codons or oligonucleotides favourable in seasonal human viruses but not in H1N1/09

	Segment 1 (Inc/Dec)	Segment 2 (Inc/Dec)	Segment 3 (Inc/Dec)	Segment 4 (Inc/Dec)	Segment 5 (Inc/Dec)	Segment 6 (Inc/Dec)	Segment 7 (Inc/Dec)	Segment 8 (Inc/Dec)
Codon	40/27	36/31	32/26	34/16	18/13	25/12	16/4	10/12
Di	25/22	26/18	26/12	40/25	19/9	25/14	14/3	9/8
Tri	33/20	32/25	33/14	47/38	22/9	28/17	16/3	17/11
Tetra	55/28	47/33	24/30	47/39	16/17	36/18	17/6	13/9
Sum	153/97	141/107	115/82	168/118	75/48	114/61	63/16	49/40

Inc/Dec: the number of increase or decrease of codons or oligonucleotides of attention.

sequence level. Notably, this tendency differed among segments. For example, the tendency was evident in segment 1 harbouring PB2 gene (Fig. 6): approximately two times more losses than gains in total (97/185: Inc/Dec in Table 2). It should be mentioned that the PB2 of the H1N1/09 was derived from an avian virus.[17,18] In the case of the segment-2-harbouring PB1 gene, which was originally derived from the human seasonal virus, the tendency was less evident and slightly reversed (Fig. 6) (147/139 in Table 2), suggesting that a large portion of the base changes observed in this gene might represent non-directional, stochastic mutation processes.

In Table 3, the number of codons or oligonucleotides favourable for seasonal human viruses but not for both H1N1/09 and avian viruses (i.e. codons or oligonucleotides predicted to increase in the near future) that were actually gained or lost in the strains isolated in the latest stage was listed (Inc or Dec in Table 3). In this case, the number of the gained codons or oligonucleotides exceeded the number of those lost (green vertical bars in Fig. 6), supporting the view that the favourable codons or oligonucleotides predicted with BLSOM were actually favourable for growth in human cellular environments. The tendency again differed among segments, and this was evident in segments 1, 6, and 7 but less evident in segment 2 (Table 3) (Fig. 6). General trend of the difference among segments was similar with that observed in Table 2, but the direction was reversed. Although one important factor contributing to the difference among segments appeared to be due to the difference in their evolutionary histories, there should exist other factors which are related to adaptation mechanisms of an invader virus in a new host. For example, the high level of losses of unfavourable codons and oligonucleotides and of gains of favourable codons and oligonucleotides in segment 1 suggests a possibility that the PB2 protein and/or RNA of the invader viruses may feel foreign in the human cellular environment because of its interaction with various host cellular factors.[25,32,33,37] Detailed inspection of differences among genes may provide information about molecular mechanisms of adaptation processes in a new host, and therefore, additional information useful for predicting sequence changes which will occur in the invader virus genome in the near future. Effects of these predicted sequence changes on protein functions, including bindings with antiviral drugs and antibodies, can be studied in detail.

Supplementary Data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Acknowledgements: We wish to thank Dr Kimihito Ito (the Research Center for Zoonosis Control, Hokkaido University) for valuable suggestions and discussions, and the editor for valuable comments. The computation was done in part at the Earth Simulator of Japan Agency for Marine-Earth Science and Technology.

Funding

This work was supported by the Integrated Database Project and 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. Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J. and Fitch, W.M. 1999, Predicting the evolution of human influenza A, *Science*, **286**, 1921–5.
2. Suzuki, Y. and Gojobori, T. 1999, A method for detecting positive selection at single amino acid sites, *Mol. Biol. Evol.*, **16**, 1315–28.
3. Igarashi, M., Ito, K., Yoshida, R., et al. 2010, Predicting the antigenic structure of the pandemic (H1N1) 2009 influenza virus hemagglutinin, *PLoS One*, **5**, e8553.
4. Nei, M. 1987, *Molecular Evolutionary Genetics*, Columbia University Press: New York.
5. Kumar, S., Nei, M., Dudley, J. and Tamura, K. 2008, MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences, *Brief Bioinform.*, **9**, 299–306.
6. Abe, T., Kanaya, S., Kinouchi, M., et al. 2003, Informatics for unveiling hidden genome signatures, *Genome Res.*, **13**, 693–702.
7. 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.
8. Kohonen, T. 1982, Self-organized formation of topologically correct feature maps, *Biol. Cybern.*, **43**, 59–69.
9. Kohonen, T., Oja, E., Simula, O., Visa, A. and Kangas, J. 1996, Engineering applications of the self-organizing map, *Proc. IEEE*, **84**, 1358–84.
10. Karlin, S., Campbell, A.M. and Mrazek, J. 1998, Comparative DNA analysis across diverse genomes, *Annu. Rev. Genet.*, **32**, 185–225.
11. 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.
12. 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.
13. Kosaka, T., Kato, S., Shimoyama, T., et al. 2008, The genome of *Pelotomaculum thermopropionicum* reveals niche-associated evolution in anaerobic microbiota, *Genome Res.*, **18**, 442–8.
14. Wilmes, P., Andersson, A.F., Lefsrud, M.G., et al. 2008, Community proteogenomics highlights microbial

strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal, *ISME J.*, **2**, 853–64.
15. Dick, G.J., Andersson, A.F., Baker, B.J., et al. 2009, Community-wide analysis of microbial genome sequence signatures, *Genome Biol.*, **10**, R85.
16. Centers for Disease Control and Prevention. 2009, Swine influenza A (H1N1) infection in two children—South California, March–April 2009, *Morb. Mortal. Wkly Rep.*, **58**, 400–2.
17. Smith, G.J., Vijaykrishna, D., Bahl, J., et al. 2009, Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic, *Nature*, **459**, 1122–5.
18. Garten, R.J., Davis, C.T., Russell, C.A., et al. 2009, Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans, *Science*, **325**, 197–201.
19. Hirahata, M., Abe, T., Tanaka, N., et al. 2006, Genome information broker for viruses, *Nucl. Acids Res.*, **35**, D339–42.
20. Bao, Y. 2008, The influenza virus resource at the National Center for Biotechnology Information, *J. Virol.*, **82**, 596–601.
21. 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.
22. García-Sastre, A. 2001, Inhibition of interferon-mediated antiviral responses by influenza A viruses and other negative-strand RNA viruses, *Virology*, **279**, 375–84.
23. Voinnet, O. 2005, Induction and suppression of RNA silencing: insights from viral infections, *Nat. Rev. Genet.*, **6**, 206–20.
24. Nelson, M.I. and Holmes, E.C. 2007, The evolution of epidemic influenza, *Nat. Rev. Genet.*, **8**, 196–205.
25. Alexey, A. and Moelling, K. 2007, Dicer is involved in protection against influenza A virus infection, *J. Gen. Virol.*, **88**, 2627–35.
26. Rabadan, R., Levine, A.J. and Robins, H. 2006, Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes, *J. Virol.*, **80**, 11887–91.
27. Wong, E.H., Smith, D.K., Rabadan, R., Peiris, M. and Poon, L.L. 2010, Codon usage bias and the evolution of influenza A viruses, *BMC Evol. Biol.*, **10**, 253.
28. Ikemura, T. 1981, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes, *J. Mol. Biol.*, **146**, 1–21.
29. Ikemura, T. 1985, Codon usage and transfer RNA content in unicellular and multicellular organisms, *Mol. Biol. Evol.*, **2**, 13–34.
30. Sharp, P.M. and Matassi, G. 1994, Codon usage and genome evolution, *Curr. Opin. Gen. Dev.*, **4**, 851–60.
31. Sueoka, N. 1995, Intrastrand parity rules of DNA base composition and usage biases of synonymous codons, *J. Mol. Evol.*, **40**, 318–25.
32. Domingo, E. and Holland, J.J. 1997, RNA virus mutations and fitness for survival, *Annu. Rev. Microbiol.*, **51**, 151–78.
33. Greenbaum, B.D., Levine, A.J., Bhanot, G. and Rabadan, R. 2008, Patterns of evolution and host gene mimicry in influenza and other RNA viruses, *PLoS Pathog.*, **4**, e1000079.
34. Itzstein, M.V. 2007, The war against influenza: discovery and development of sialidase inhibitors, *Nat. Rev. Drug Discov.*, **6**, 967–74.
35. Le, Q.M., Kiso, M., Someya, K., et al. 2005, Avian flu: isolation of drug-resistant H5N1 virus, *Nature*, **437**, 1108.
36. Fodor, E., Devenish, L., Engelhardt, O.G., et al. 1999, Rescue of influenza A virus from recombinant DNA, *J. Virol.*, **73**, 9679–82.
37. Yamada, S., Hatta, M., Staker, B. L., et al.2010, Biological and structural characterization of a host-adapting amino acid in influenza virus, *PLoS Pathog.*, **6**, e1001034.