EXPERIMENTAL and MOLECULAR MEDICINE, Vol. 38, No. 6, 643-651, December 2006

Epidemiological comparisons of codon usage patterns among HIV-1 isolates from Asia, Europe, Africa and the Americas

Insung Ahn[1,2] and Hyeon Seok Son[2,3]

[1]Bioinformatics Team, Supercomputing Center
Korea Institute of Science and Technology Information
Daejeon 305-806, Korea
[2]Laboratory of Computational Biology and Bioinformatics
Institute of Health and Environment
Graduate School of Public Health
Seoul National University
Seoul 110-799, Korea
[3]Corresponding author: Tel, 82-2-740-8864;
Fax, 82-2-762-9105; E-mail, hss2003@snu.ac.kr

Accepted 20 October 2006

Abbreviations: CRF, circulating recombinant form; RSCU, relative synonymous codon usage

Abstract

To investigate the genomic properties of HIV-1, we collected 3,081 sequences from the HIV Sequence Database. The sequences were categorized according to sampling region, country, year, subtype, gene name, and sequence and were saved in a database constructed for this study. The relative synonymous codon usage (RSCU) values of *matrix*, *capsid*, and *gp120* and *gp41* genes were calculated using correspondence analysis. The synonymous codon usage patterns based on the geographical regions of African countries showed broad distributions; when all the other regions, including Asia, Europe, and the Americas, were taken into account, the Asian countries tended to be divided into two groups. The sequences were clustered into nine non-CRF subtypes. Among these, subtype C showed the most distinct codon usage pattern. To determine why the codon usage patterns in Asian countries were divided into two groups for four target genes, the sequences of the isolates from the Asian countries were analyzed. As a result, the synonymous codon usage patterns among Asian countries were divided into two groups, the southern Asian countries and the other Asian countries, with subtype 01_AE being the most dominant subtype in southern Asia. In summary, the synonymous codon usage patterns among the individual HIV-1 subtypes reflect genetic variations, and this bioinformatics technique may be useful in conjunction with phylogenetic methods for predicting the evolutionary patterns of pandemic viruses.

Keywords: computational biology; genes, env; genes, gag; genomics; HIV-1; sequence analysis, RNA

Introduction

Over the past 30 yr, human immunodeficiency virus type 1 (HIV-1) infection has increased and now threatens the health of people worldwide. HIV-1, which is the most common cause of AIDS, has already infected more than 50 million people. An estimated 3.2 million individuals in Asia are HIV-infected, with 1.1 million people becoming infected in 2005 alone (*AIDS* Epidemic Update 1). China had an estimated 70,000 new HIV infections in 2005, according to reports by the Chinese Ministry of Health, the Joint United Nations Program on HIV/AIDS (UNAIDS), and the World Health Organization (*AIDS* Epidemic Update 2). Most HIV infections (approximately 80%) are related to intravenous drug use and unprotected sexual intercourse. Furthermore, the number of people with HIV in North America and Western and Central Europe rose to 1.9 million in 2005, with approximately 65,000 people having acquired HIV in the past year (*AIDS* Epidemic Update 1).

The strains of HIV-1 can be classified into three groups, group M (for main), group N (for non-M/non-O or new), and group O (for outlier), which are the result of three cross-species transmissions of chimpanzee viruses into the human population (Pierre *et al.,* 1994; Beatrice *et al.*, 2000). Group O HIV-1 is usually found in west-central Africa, and group N, which was discovered in 1998 in Cameroon, rarely appears. Most of the HIV-1 infections belong to HIV-1 group M, which is divided into nine distinct subtypes (A, B, C, D, F, G, H, J and K) and 14 intersubtypes of HIV-1 recombinants, which are known as circulating recombinant forms (CRFs). Each subtype can be classified according to the criteria listed below. The subtypes are approximately equidistant from one another in terms of the *env* gene, and the *env* phylogenetic tree is, for the most part, congruent with the *gag* phylogenetic tree. Moreover, two or more samples are required to define a new sequence subtype. Envelope-coding sequences were

used as the basis for classification in the 1992 compendium but, as the database of gene sequences has expanded, this basis has been increasingly challenged. For example, although subtype E viruses are equidistant from viruses of subtypes A to D in terms of their *env*-coding sequences, they appear to be similar to subtype A viruses with respect to their *gag*-coding sequences. The *gag*-coding sequences of subtype E viruses have not yet been documented. The subtypes have been used successfully as molecular epidemiological markers to track the course of the HIV-1 pandemic. For instance, Africa has all the known HIV-1 subtypes and groups, reflecting the African origin of the epidemic (Wouter *et al.,* 1997). Subtype B viruses are prevalent in the Americas and Western Europe, and subtype C is the most prevalent HIV type worldwide (José and Natt, 2000).

Synonymous codon usage pattern analysis is a common method for examining the origin of species in many fields (Bulmer, 1978; Shields and Sharp, 1987; Moriyama and Hartl, 1993; Stenico *et al.*, 1994; Mclnerney, 1998; Kanaya *et al.*, 2001; Duret, 2002; Lynn *et al.*, 2002). Synonymous codons often encode common amino acids during protein synthesis. However, they are not used randomly, and some codons are used more frequently than others (Moriyama and Hartl, 1993; Mclnerney, 1998; Duret, 2002; Lynn *et al.*, 2002). In prokaryotes, such as thermophilic bacteria, the codon usage in highly expressed genes shows a shift towards a more restricted set of preferred synonymous codons (Lynn *et al*., 2002). Codon usage patterns tend to mirror the distribution of tRNA abundances (Shields and Sharp, 1987; Stenico *et al.*, 1994). The relative synonymous codon usage (RSCU) values may be virus-specific (Gu *et al.*, 2004) and may not be affected by translational selection or gene length (Jenkins and Holmes, 2003). Jenkins and Holmes (2003) analyzed codon usage patterns by applying correspondence analysis, which is a type of statistical perceptual mapping that relates categories of a contingency table. Correspondence analysis is a compositional technique in which the perceptual map is based on the association between objects and a set of specified descriptive characteristics or attributes (Hair *et al.*, 1998; Johnson and Wichern, 2002). It should be noted that almost all of the codon usage studies using correspondence analysis have been performed in bacteria or higher species but not in viruses.

This study investigated the characteristics of genomic patterns, particularly with respect to the codon usage patterns of HIV-1, on the basis of continental regions, such as Asia, Europe, Africa, and the Americas, as well as nine non-CRF subtypes. The HIV-1 genome contains three major regions, *gag*, *pol*, and *env*. We chose to determine the subtypes of the *gag* and *env* genes, which encode the inner and outer structures of HIV-1, respectively. The protease, reverse transcriptase, and integrase enzymes derived from the *pol* gene as well as *tat* (= transcriptional activation *via* RNA target) gene (Lee *et al*., 2004) were excluded, as the focus was on genes that are related to the specific structure of HIV-1. The *matrix*, which is located under the viral plasma membrane, plays an important role in transporting the *gag-pol* precursor polyprotein to the plasma membrane, while the capsid, which is derived from the *gag* gene, like the *matrix*, constitutes the core shell of the viral particle. If a mutation occurs in the capsid protein, that virus is not coated during the assembling process. Among the *env*-encoded proteins, the surface glycoprotein *gp120* triggers the viral entry process, while *gp41* mediates the fusion interactions between the viral and cellular membranes in the initial stage of infection. These four genes, which are encoded from the *env* and *gag* gene regions, were analyzed in the present study.

Materials and Methods

Nucleotide sequences

HIV-1 sequence files in the FASTA format were collected from the HIV Sequence Database (http://hiv.lanl.gov), which is based on HIV sequences downloaded from GenBank (http://www.ncbi.nih.gov/Genbank), and all of the sequences were divided into four gene groups: *matrix* (784 sequences), *capsid* (829 sequences), *gp120* (715 sequences), and *gp41* (753 sequences), using Java and the MySQL database (version 3.23) on the Linux operating system. The accession number was a primary key, and other fields, such as sampling region, country, year, subtype, gene name, and sequence, were also created in the database, which was constructed for this study. The sampling years of the sequences were from 1991 to 2004, and these sequences were grouped into five geographical regions, which included Asia, Europe, Africa, and North and South America, on the basis of the 'sampling region' field. Any sequences that had ambiguous characters or length were pre-screened, and no redundant sequences were sampled from the same patient. The non-CRF (circulating recombinant form) subtypes, which included subtypes A, B, C, D, F, G, H, J and K, were extracted from the database and used to analyze the codon usage patterns of different geographical regions.

Relative synonymous codon usage (RSCU)

In correspondence analysis, the RSCU value for

each codon is usually used to prevent the amino acid composition from influencing the codon usage values for each gene (Sharp and Li, 1986). The RSCU value is the number of times a particular codon is observed relative to the number of times the codon would be observed in the absence of any codon usage bias. If codon usage bias is absent, the RSCU value is 1.00. The RSCU was calculated as:

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$$

where X_{ij} is the frequency of occurrence of the *j*th codon for the *i*th amino acid, and n_i is the number of codons for the *i*th amino acid. For the correspondence analysis, each gene was represented as a 59–dimensional vector, with the exclusion of start and stop codons and the UGG codon, which encodes tryptophan without synonymous codons. All 59 codons were then pooled and collated in a contingency table according to the species genome in which they were included. In this study, we used correspondence analysis to compare the differences in RSCU values among HIV-1 isolates on the basis of geographical region, non-CRF subtype or country of collection.

Correspondence analysis

Correspondence analysis is a type of multivariate analysis that represents associations either as tabulated frequencies or graphically as counts. Each gene is presented as a 59-dimensional vector, and

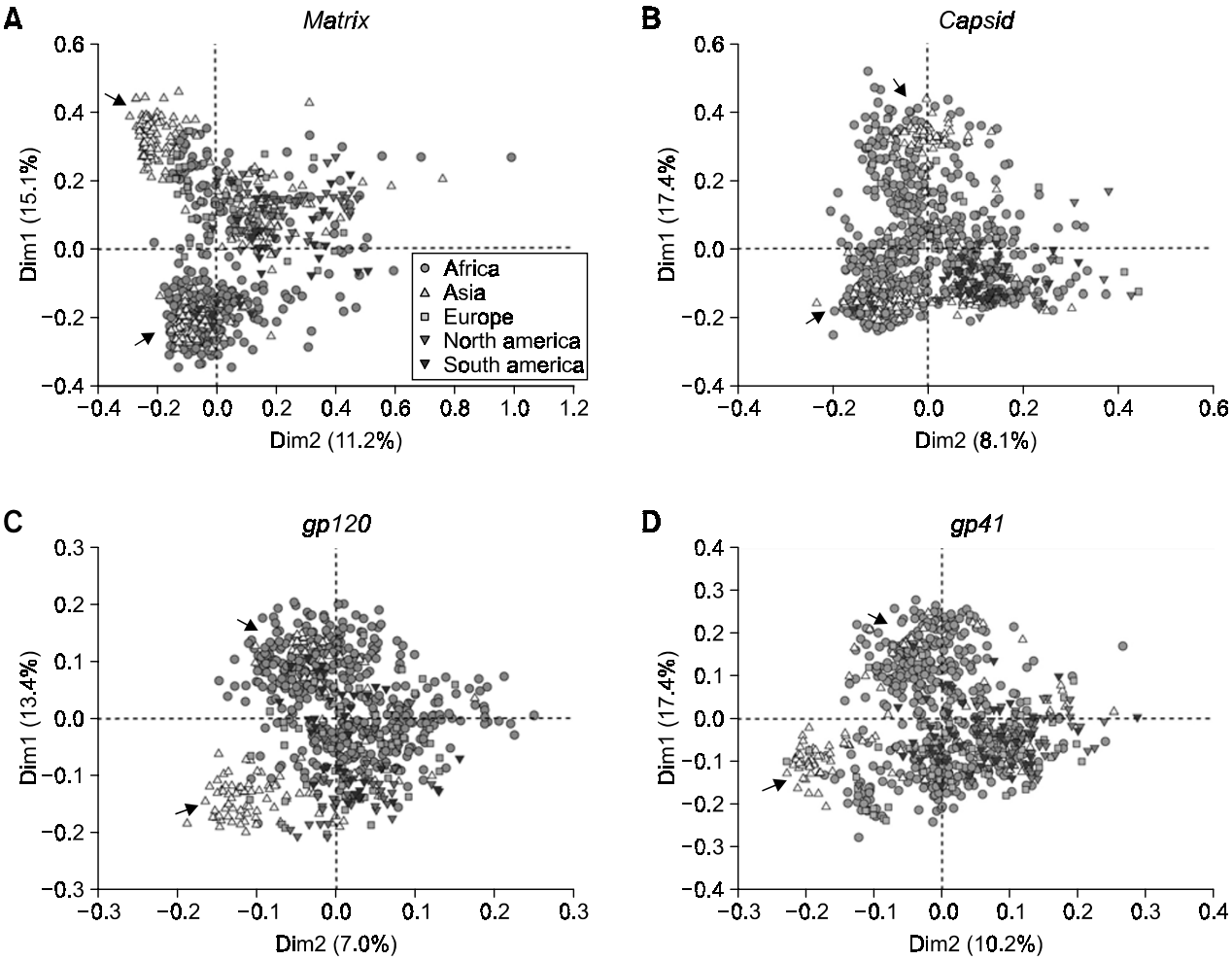

Figure 1. A plot of the values of the first and second axes from the correspondence analysis values for the (A) *matrix*, (B) *capsid*, (C) *gp120* and (D) *gp41* genes of HIV-1 isolates from Asia, Europe, Africa, and North and South America. Dim1 and Dim2 represent the values of the first and second dimensional factors, respectively, for each sequence. The percentage in each parenthesis denotes the percentage inertia of that axis in the correspondence analysis. Each distinct group of Asian isolates (triangles filled in yellow) is marked with a short arrow.

each dimension corresponds to the RSCU value of one sense codon, excluding AUG, UGG, and three stop codons (Gu *et al.*, 2004). For a contingency table with *I* rows and *J* columns, the plot produced by correspondence analysis contains two sets of points: one set of *I* points that corresponds to the rows, and one set of *J* points that corresponds to the columns. The positions of the points reflect the associations. The RSCU values for 59 codons, as described above, were calculated, and correspondence analysis was performed using the SAS statistical program version 9.1 (Cary, 2004).

The usual output from a correspondence analysis includes the "best" two-dimensional representation of the data, along with the coordinates of the plotted points and a measure (called the inertia) of the amount of information retained in each dimension (Johnson and Wichern, 2002). The results provide information similar to that produced by factor analysis techniques, and they allow exploration of the structure of the categorical variables included in the table. Genes or species that are strongly associated, as measured by their chi-square distances, lie in a similar direction from the origin (Johnson and Wichern, 2002; Perrière and Thioulouse, 2002). The chi-square distance between two coordinates with the same row or column value, called the Euclidean distance (Perrière and Thioulouse, 2002; Gu *et al.*, 2004), has important statistical meaning, whereas there is no significant statistical meaning attached to the relationship between the row coordinate and column coordinate. The SigmaPlot 8.0 program was used to create graphical correspondence analysis plots using coordinates that consisted of the first-

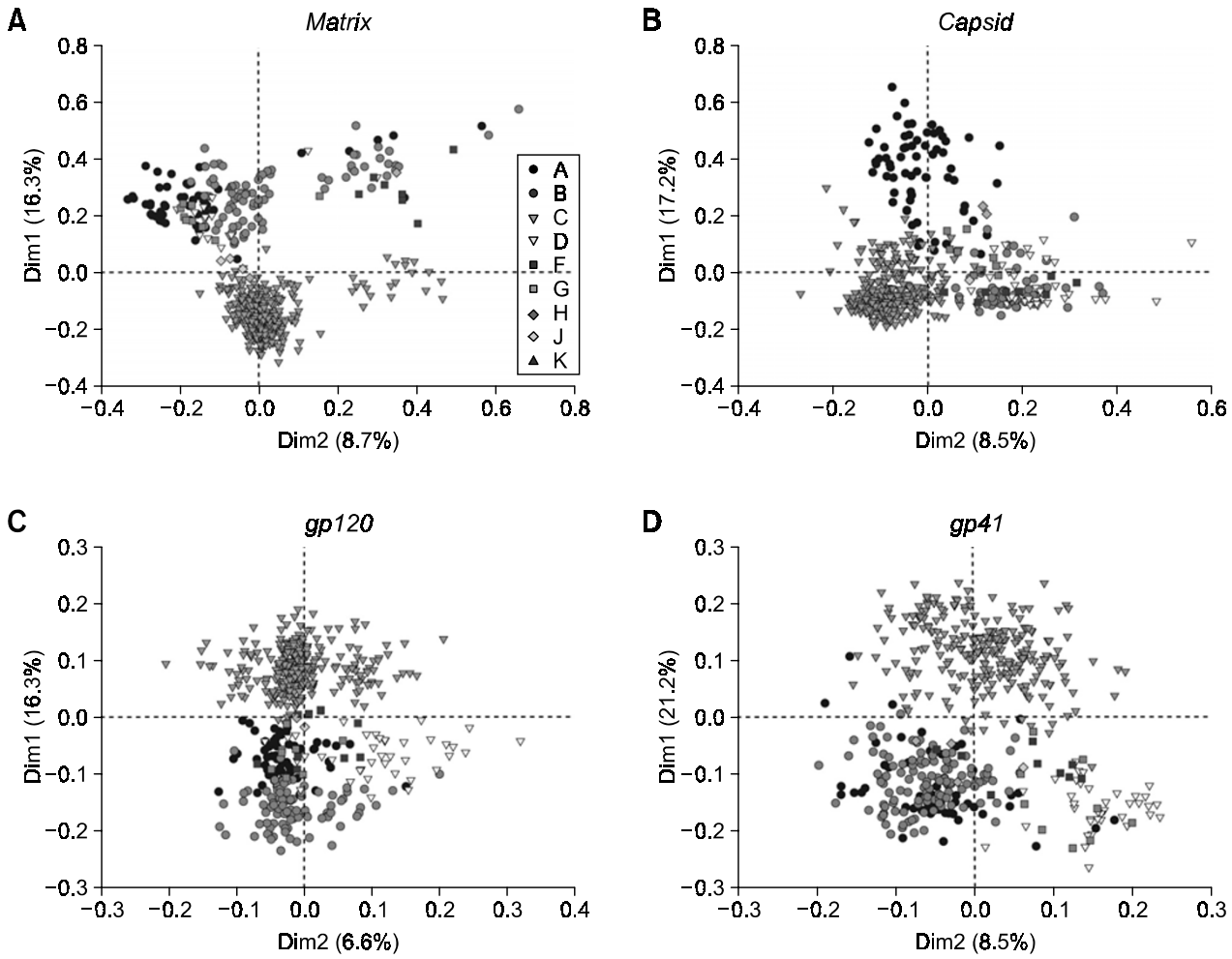

Figure 2. A plot of the values of the first and second axes from the correspondence analysis values for the (A) *matrix*, (B) *capsid*, (C) *gp120* and (D) *gp41* genes of the non-CRF HIV-1 subtypes from Asia, Europe, Africa, and North and South America. Dim1 and Dim2 represent the values of the first and the second dimensional factors, respectively, for each sequence. The percentage in each parenthesis denotes the percentage inertia of that axis in the correspondence analysis.

and second-dimensional factors that resulted from the correspondence analysis.

Results

Synonymous codon usage patterns of HIV-1 isolates from Asia, Europe, Africa, and America

To compare the overall synonymous codon usage patterns, we created an analytical database that included the 3,081 sequences, including the genes from the *matrix*, *capsid*, *gp120*, and *gp41*, and performed correspondence analysis with the RSCU values for each sequence using Java codes (Figure 1). Of the four genes, two *gag*-originating genes, for the *matrix* and *capsid*, showed more biased synonymous codon usage patterns for the Dim1 gene than for the two genes that originated from the *env* gene region. The unit ranges of Dim1 for the *matrix* and *capsid* were 0.69 (-0.34～0.35) and 0.68 (-0.25～0.43), while those for *gp120* and *gp41* were 0.41 (0.21～0.20) and 0.56 (-0.28～0.28), respectively. Isolates from African countries revealed broad codon usage distributions; taking all the other geographical regions into consideration, Asian countries tended to divide the isolates into two groups on the basis of Dim1. European isolates also showed relatively broad distributions but the differences in the ranges were narrower than those for the African isolates. Isolates from North and South America were clustered together for the *matrix*, *capsid*, and *gp41* genes, but were divided into two groups for *gp120*.

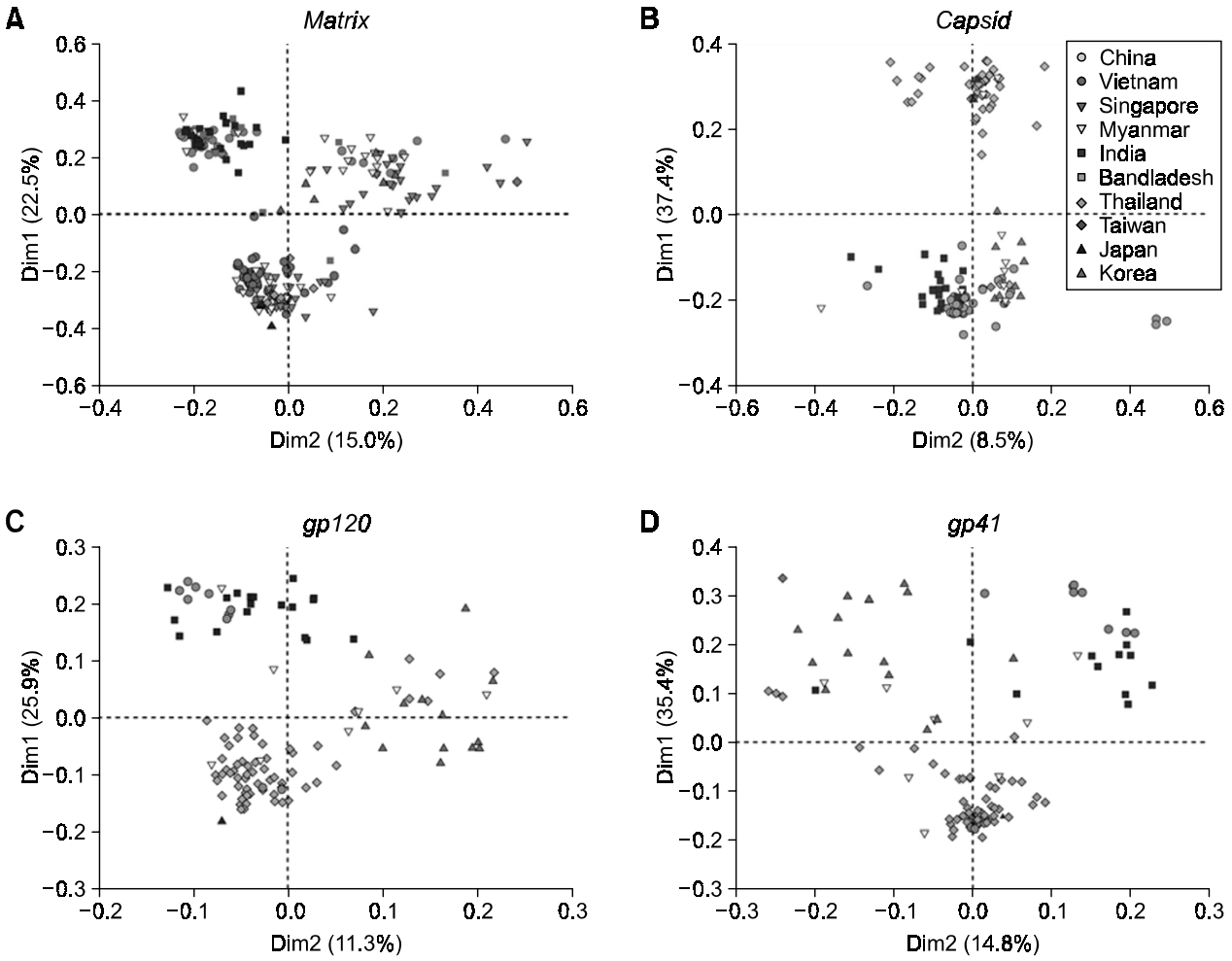

Figure 3. A plot of the values of the first and second axes from the correspondence analysis values for the (A) *matrix*, (B) *capsid*, (C) *gp120* and (D) *gp41* genes of HIV-1 isolates from Asian countries. Dim1 and Dim2 represent the values of the first and the second dimensional factors, respectively, for each sequence. The percentage in each parenthesis denotes the percentage inertia of that axis in the correspondence analysis.

Synonymous codon usage patterns of HIV-1 isolated among nine non-CRF subtypes

In addition to the codon usage analysis using the sequence datasets that were divided according to geographical region, we performed correspondence analysis using the RSCU values for each non-CRF subtype, i.e., subtypes A, B, C, D, F, G, H, J, and K, from all the countries available (Figure 2). In contrast to the result described above (Figure 1), the HIV-1 subtypes were readily classified on the basis of their synonymous codon usage patterns. Among these subtypes, subtype C showed the most distinct distribution. With regard to Dim1 in the correspondence analysis plots, subtype C viruses were located opposite the other subtypes, with the exception of the *capsid* gene. On the other hand, subtype A viruses showed distributions for the *capsid* genes that were distinct from those of the other subtypes, although they shared similar patterns with subtype B for the other genes. Subtype D showed similar patterns to subtype B for both the *matrix* and *capsid* genes (Figure 2A and B), but revealed somewhat different distributions for *gp120* and *gp41* (Figure 2C, D).

The different codon usage patterns among Asian HIV-1 isolates

To understand why the codon usage distributions of isolates from the Asian region were divided into two distinct groups (Figure 1), we extracted the sequences that originated in the Asian countries, and analyzed them on the basis of individual countries of origin (Figure 3). For the *capsid*, *gp120*, and *gp41* genes, the HIV-1 sequences from Thailand commonly showed synonymous codon usage patterns that

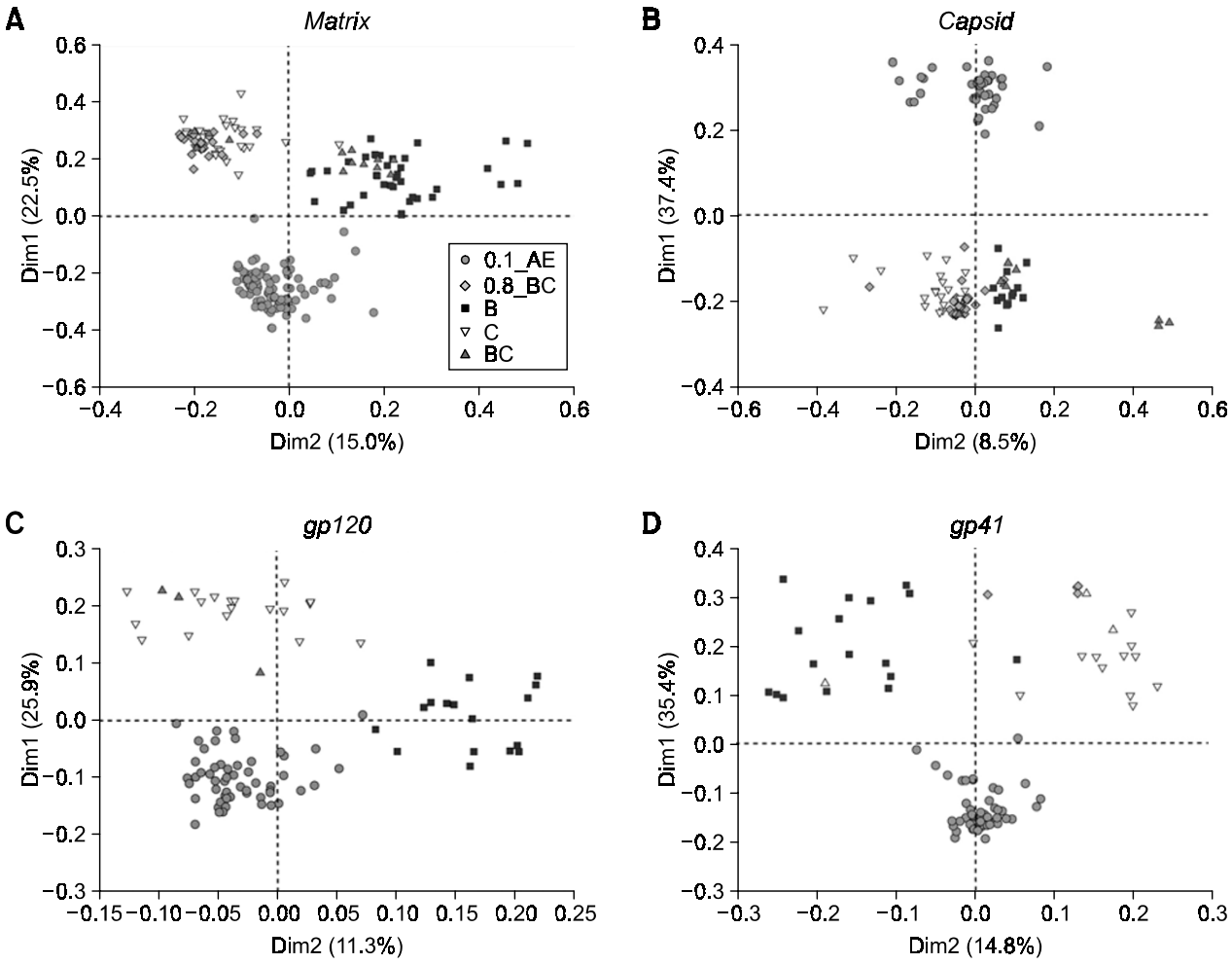

Figure 4. A plot of the values of the first and second axes from the correspondence analysis values for the (A) *matrix*, (B) *capsid*, (C) *gp120* and (D) *gp41* genes of the five major HIV-1 subtypes from Asian countries. Dim1 and Dim2 represent the values of the first and the second dimensional factors, respectively, for each sequence. The percentage in each parenthesis denotes the percentage inertia of that axis in the correspondence analysis.

were distinct from those of other countries (Figure 3B, C and D), and the sequences from Vietnam, Myanmar, Singapore, Japan, and Thailand for the matrix genes were clustered as a single group in the corresponding analysis plots (Figure 3A). Some sequences from Singapore and Myanmar were located at the same position as those from Korea and China. Using the same sequences, we classified the plots on the basis of their major subtypes, which included non-CRF and CRF subtypes, to determine the relationships between HIV-1 subtypes and their distributions in Asian countries (Figure 4). The five major HIV-1 subtypes identified in this database, but not in the populations, were 01_AE, 08_BC, BC, B, and C. When compared with the results shown in Figure 3, it appears that high levels of 01_AE subtype viruses occur in southern Asian countries, which include Vietnam, Singapore, Thailand, and Myanmar. Unlike the sequences from Korea, the sequences from Japan were colocalized with this southern Asian group. Some of the sequences from Myanmar were also found in the other groups, which may reflect the geographical location of Myanmar, which lies between the southern Asian countries and the Chinese continent. HIV-1 isolates from China and India tended to be grouped together; subtypes BC and C were dominantly distributed in these countries, while subtype 01_AE viruses were mainly distributed in southern Asian countries. Subtype B viruses were predominate in Korea.

Discussion

Large amounts of genomic data have been collected and distributed on the Web by various organizations such as the U.S. National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI). Thanks to these public databases, not only the survey data but also the genomic data have become important resources for the field of genetic epidemiology. Since the initial reports of HIV-1 virus in Africa, it has spread worldwide and generated biological subtypes, such as the nine non-CRF viruses and several other CRFs. As a result of geographical constraints, the distributions of these subtypes are so uneven that the differences in HIV-1 subtypes between countries can be used as a marker of epidemiological diversification. In this study, we determined the codon usage differences between HIV-1 isolate from each geographical region, to elucidate their evolutionary patterns.

To investigate the overall patterns of synonymous codon usages among the isolates from Asia, Europe, Africa, and North and South America, we extracted 3,081 nucleotide sequences from the HIV Sequence Database (Figure 1). Isolates from African countries showed broad codon usage distributions, taking all the other geographical regions into consideration. This result appeared reasonable because almost all HIV-1 subtypes are still distributed in African countries, especially near the central African region where the first HIV-1 infection occurred. In concordance with the previous phylogenetic analyses of HIV-1 subtypes (Wouter *et al.*, 1997; Nicole *et al.*, 2000; Feng *et al.*, 2001), our results show that all the synonymous codon usage patterns exist in African countries. Unlike the African isolates, the isolates from Asian countries were divided into two distinct groups, particularly for the matrix and capsid genes. This is an interesting result, as the different codon usage patterns for these two groups may provide some clues as to the differentiation of epidemiological patterns of HIV-1 subtypes. We conducted an additional analysis that focused on the Asian isolates. On the other hand, the broader unit ranges observed for Dim1 of the *gag*-origin genes, such as the *matrix* and *capsid* genes, as compared to those of *gp120* and *gp41*, suggest that genes responsible for virus-host interactions are more likely to be less biased or varied in terms of synonymous codon usage patterns than other structural genes. As the envelope glycoproteins *gp120* and *gp41* perform important roles in receptor binding and in interacting with neutralizing antibodies, they appear to be more conserved than other structural genes (Lee *et al.*, 2006).

The differences in synonymous codon usage patterns among the subtypes were also analyzed using the nine non-CRF subtypes. We excluded the CRF subtypes from this study because the focus was on the unique, not mixed, codon usage patterns of each subtype (Figure 2). All the sequences of the nine non-CRF subtypes were extracted from the database and analyzed. Interestingly, HIV-1 subtypes revealed a clear classification with respect to their codon usage patterns, regardless of the sampling country. Among these non-CRF subtypes, subtype C showed the most distinct distribution pattern. Subtype C viruses, initially a rarely reported subtype, now dominate the global pandemic and constitute the majority of the newly transmitted cases of HIV-1 in many developing countries. There is growing evidence to suggest that subtype C viruses display characteristics that distinguish them from other subtypes, and that these differences affect transmission and pathogenesis (Hong *et al.*, 2002). In the present study, this subtype proved to have different synonymous codon usage patterns. HIV-1 subtype C, which is currently ranked as the third most dominant HIV-1 subtype in the world, appears to have been successfully transmitted from the original source in central

Africa to countries such as India and Saudi Arabia (HIV-1 Geography Maps, 2006).

In the analysis of the overall patterns of synonymous codon usage among isolates from Asia, Europe, Africa, and North and South America, the Asian isolates tended to separate into two groups on the basis of Dim1. To understand this phenomenon, we extracted the sequences for the Asian isolates and analyzed the RSCU patterns using correspondence analysis (Figures 3 and 4). Among the Asian countries, all the HIV-1 subtypes were distributed unevenly, showing distinct synonymous codon usage patterns for different geographical regions. Interestingly, the sequences from southern Asian countries, which include Vietnam, Singapore, and Thailand, were grouped together, and the CRF subtype 01_AE viruses predominated. Although the dataset from Japan used in this study was small, the CRF 01_AE subtype viruses were also found in Japan, and they showed the same codon usage patterns as isolates from the other southern Asian countries. When we compared the patterns of the nine non-CRF subtypes (Figure 2), subtype C showed the most distinct patterns. However, the CRF subtype 01_AE isolates weakened the effect of subtype C, being located in a distinctly different region to the other subtypes, including subtype C. According to Nicole *et al.* (2000), the highest prevalence of CRF 01_AE is in the north of the Democratic Republic of Congo (DCR), which is known as the epicenter for HIV-1 group M viruses in Africa; this recombinant subtype uses the synonymous codons in a different manner in protein synthesis than do the other subtypes in our study. For various reasons, the CRF 01_AE subtype seems to have adapted successfully to the southern Asian countries, including Japan in the Far East. In contrast, subtype 01_AE and subtypes BC and C predominate in China and India, and cluster separately in the correspondence plots. To date, subtype C viruses have been found to play an important role only in the epidemics of southern Africa (Wouter *et al.*, 1997). However, subtype C viruses are found with increasing frequency in India, which is the gateway to the Asian continent from the African continent. According to population-based studies, subtype B is the most predominant subtype in the world (HIV-1 Geography Maps, 2006), especially in Europe and the United States (Ras *et al.*, 1983). Unlike subtypes 08_BC, BC, and C, subtype B was distinctly grouped (Figure 4), showing similar but not identical codon usage patterns to other subtypes, with the exception of subtype 01_AE. Some sequences from Singapore and Myanmar colocalized with those from Korea and China. Since Myanmar is located between the southern Asian countries and China, it is reasonable to find intermediate patterns between those two regions. On the other hand, for isolated Singapore, it is difficult to explain why the two subtypes of subtype B subtype 01_AE were dominant. Subtype B is often found in Europe and America, so tourists from these countries to Singapore may be transmitters of subtype B.

For many years it was believed that errors in reverse transcriptase and the presence of virion RNA in dimer form were the driving forces behind the genetic variation of HIV. However, since 1995, many studies have suggested that recombination plays a more important role in the genetic diversification of HIV-1 than was previously thought (Wouter *et al.*, 1997). In the present study, we show that each HIV-1 subtype uses synonymous codons in a specific way, and that these codon usage patterns may play a role in the distribution of HIV-1 subtypes. In summary, the synonymous codon usage patterns among the HIV-1 subtypes reflect genetic variability, and this bioinformatics technique in conjunction with phylogenetic methods may be useful in predicting the evolutionary patterns of pandemic viruses. This study extends our knowledge of HIV-1 subtypes in terms of genomic analysis. Further genetic epidemiology studies using genomic patterning on a population basis are needed to substantiate these findings, and bioinformatics techniques may be valuable tools in this respect.

Acknowledgment

We acknowledge the contribution of all those who have made their invaluable data publicly available. This work was supported by the Brain Korea 21 Project (2006).

References

AIDS Epidemic Update 1 from United Nations Programme on HIV/AIDS (UNAIDS). HIV Data: Regional Information, January 2006. Available at: www.unaids.org

AIDS Epidemic Update 2 from United Nations Programme on HIV/AIDS (UNAIDS). New data show growing AIDS epidemic in China, January 2006. Available at: http://data.unaids.org/Media/Press-Releases03/PR_china_060125_en.pdf

Beatrice HH, George MS, Kevin MDC, Paul MS. AIDS as a zoonosis: scientific and public health implications. Sci 2000; 287:607-14

Bulmer MA. Statistical analysis of nucleotide sequences of introns and exons in human genes. Mol Biol Evol 1987;l4: 395-405

Cary NC. *SAS® R9.1.2 Qualification Tools User's Guide,* 1th Ed, 2004, SAS Institute Inc., NC, USA

Duret L. Evolution of synonymous codon usage in metazoans. Curr Opin Gene Dev 2002;12:640-9

Feng G, Nicole V, Yingying L, Stanley AT, Yalu C, Leondios GK, David DH, Jinwook K, Myoungdon O, Kangwon C, Mika S, David LR, George MS, Beatrice HH, Martine P. Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation. AIDS Res Hum Ret 2001;17:675-88

Gu W, Zhou T, Ma J, Sun X, Lu Z. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales. Virus Res 2004;101:155-61

Hair JFJr, Anderson RE, Tatham RL. *Multivariate Data Analysis*, 5th Ed, 1998, Prentice-Hall International, Inc., NJ, USA

HIV-1 Geography maps at Los Alamos HIV Sequence Database. April 2006. Available at: http://www.hiv.lanl.gov/content/hiv-db/geography/index.html

Hong Z, Guillermo O, Qiujian D, Jun H, Chipepo K, Ganapati B, Charles W. Phylogenetic and phenotypic analysis of HIV type 1 env gp120 in cases of subtype C mother to child transmission. AIDS Res Hum Ret 2002;18:1415-23

Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 2003;92:1-7

Johnson RA. Wichern DW. *Applied Multivariate Statistical Analysis*, 5th Ed, 2002, Prentice-Hall, Inc., NJ, USA

José E, Natth B. Accelerating the development and future availability of HIV-1 vaccines: why, when, where, and now? Lancet 2000;355:2061-6

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. Codon usage and trna genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol 2001;53:290-8

Lee HJ, Ryu JY, Kim KA, Lee KS, Lee JY, Park JB, Park JS, Choi SY. Transduction of yeast cytosine deaminase mediated by HIV-1 Tat basic domain into tumor cells induces chemosensitivity to 5-fluorocytosine. Exp Mol Med 2004;36:43-51

Lee MK, Kim HK, Lee TY, Hahm KS, Kim KL. Structure-activity relationships of anti-HIV-1 peptides with disulfide linkage between D- and L-cystein at positions i and i+3, respectively, derived from HIV-1 gp41 C-peptide. Exp Mol Med 2006;38:18-26

Lynn DJ, Singer GAC, Hickey DA. Synonymous codon usage in subject to selection in thermophilic bacteria. Nucleic Acids Res 2002;30:4272-7

Mclnerney JO. Codon Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc. Natl Acad Sci 1998;95:10698-703

Moriyama EN, Hartl DL. Codon usage bias and base composition of nuclear genes in drosophila. Genetics. 1993;134: 847-58.

Nicole V, Martine P, Claire MK, Nzila N, David R, Wantabala I, Hurogo S, Kazadi T, Beni B, Eric D. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the democratic republic of Congo suggests that the HIV-1 pandemic originated in central Africa. J Virol 2000;74:10498-507

Perrière G, Thioulouse J. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res 2002; 30:4548-55

Pierre C, Andrew MB, Caroline Q, Denise G, Sophie C, Jacques C, Gérard R, Luc M, François C. Isolation and envelope sequence of a highly divergent HIV-1 isolate: definition of a new HIV-1 group. Virol 1994;205:247-53

Ras GJ, Simson IW, Anderson W, Prozesky OW, Hamerma T. Aquired immunodeficiency syndrome. A report of 2 South African cases. S Afr Med J 1983;64:140-2

Sharp PM, Li WH. Codon usage in regulatory genes in *Escherichia coli* does not reflect for 'rare' codons. Nucleic Acids Res 1986;14:7737-49

Shields DC, Sharp PM. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res 1987;15:8023-40

Stenico M, Lloyd AT, Sharp PM. Codon usage in Caenorhabditis elegans: delineation of translational selection and mutational biases. Nucleic Acids Res 1994;22:2437-46

Wouter J, Anne B, John NN. The puzzle of HIV-1 subtypes in Africa. AIDS 1997;11:705-12