ACTA ABP BIOCHIMICA POLONICA

Vol. 61, No 3/2014
441–451
on-line at: www.actabp.pl

Regular paper

Mapping of the influenza A hemagglutinin serotypes evolution by the ISSCOR method

Jan P. Radomski[1,4]✉, Piotr P. Słonimski[2*], Włodzimierz Zagórski-Ostoja[3] and Piotr Borowicz[4]

[1]Interdisciplinary Center for Mathematical and Computational Modeling, Warsaw University, Warsaw, Poland; [2]Centre de Génétique Moléculaire du CNRS & Université Pierre-et-Marie Curie, Paris, Gif–sur–Yvette, France; [3]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland; [4]Institute of Biotechnology and Antibiotics, Warsaw, Poland

Analyses and visualizations by the ISSCOR method of influenza virus hemagglutinin genes of different A-subtypes revealed some rather striking temporal relationships between groups of individual gene subsets. Based on these findings we consider application of the ISSCOR-PCA method for analyses of large sets of homologous genes to be a worthwhile addition to a toolbox of genomics — allowing for a rapid diagnostics of trends, and ultimately even aiding an early warning of newly emerging epidemiological threats.

Key words: ISSCOR descriptors; phylogenetic analysis; influenza virus; hemagglutinin; phylogenetic maps

Received: 28 May, 2014; revised: 26 August, 2014; accepted: 11 September, 2014; available on-line: 12 September, 2014

INTRODUCTION

Living organisms have very often quite biased preferences for some synonymous codons coding for the same amino acids. These differences and their variation have been extensively studied, however, no governing rules have yet been discovered. Frequencies of codons for many species are in close correlation with their genome's GC contents, but the underlying forces governing this are not clear — it might be possible, that it is the GC content which is determining a genome's amino acids predilection for the specific codons being used (Knight *et al.,* 2001). On the other hand, it might be that reverse causative relationships are in operation: codons-specific amino acids usage is a driving factor for observed GC contents. Possible factors and forces driving synonymous codons usage postulated so far include, among many others: translational optimization (Kimura, 1962; Berg & Kurland, 1997; Carbone *et al.*, 2005; Dreies *et al.,* 2004; Novozhilov *et al.,* 2007), mRNA structural effects (Zama, 1990), protein composition (Collins, 1993), and protein structure (Adzubei *et al.,* 1996), gene expression levels (Sharp & Matassi, 1994), the tRNA abundance differences between different genomes, and tRNA optimization (Ikemura, 1985; Buchan *et al.,* 2006; Rodnina *et al.,* 2001), or different mutation rates and patterns (Suoeka, 1992). Also, some other possibilities were hypothesized, like local compositional bias (Antezana & Kreitman, 1999), and even gene lengths might play a role too (Eyre-Walker, 1993).

It is clear, that many interesting biological mechanisms underlie the basic phenomenon of genetic code degeneracy. One of its aspects, however, has not been studied until recently — the question dealing with the sequential order of occurrence of synonymous codons. Obviously, an order of elements in a linear set is a different property than the frequency of elements in the set. The amino acid composition of a protein carries much less information than the amino acid sequence of such a protein, which in turn is less information intensive than a corresponding nucleotide sequence coding the same protein. This question can be formulated more precisely if we consider a given frequency of synonymous codon usage characteristic for a gene. There is a large number of different orders in which the synonymous codons can appear sequentially along the gene without changing either the amino acid sequence of the encoded protein, or the codon usage of the gene.

The genome of influenza A viruses consists of eight RNA segments that code for 10 viral proteins. Based on the antigenic specificities of the hemagglutinin (HA), or neuraminidase (NA) proteins the influenza A viruses have been divided respectively into 17 HA (H1-H17), and ten neuraminidase (N1-N10) subtypes. Accumulation of mutations in the antigenic sites of the HA and NA that alters viral antigenicity, is called the "antigenic drift". In circulating influenza viruses this antigenic drift is a major process accumulating mutations at the antibody binding sites of receptor proteins, and enabling the virus to evade recognition by hosts' antibodies. The HA protein consists of two domains, HA1 and HA2 — the HA1 domain, the major antigenic protein of influenza A viruses, contains the large majority of the antigenic sites of HA and is under constant immune-driven selection. The segmented nature of influenza genome allows also for exchange of gene segments — a process of genetic reassortment, involving type A influenza viruses of different subtypes, and may result in the so called "antigenic shift", which occurs when progeny viruses that possess a novel HA, or a novel HA and NA, emerge (Shih *et al.,* 2007; McHardy & Adams, 2009). Ahn and Son (2012) investigated genetic variations in eight major genes of the influenza A virus H3N2 serotype, and determined the evolutionary pattern present in codon bias

*Piotr P. Słonimski, passed away on April 25th, 2009, much too early, leaving behind a large portfolio of many joint genomic projects and ideas at various stages of development, some of which need long time to be fulfilled. Specifically, in the case of large assemblies of the HA orthologs, which we have started to analyze already back in 2006–2007, the sufficiently detailed data are available only recently, although the first basic results showing the peculiar triangular distributions presented here, were obtained in the 2008.

✉e-mail: janr@icm.edu.pl

Abbreviations: HA, hemagglutinin; ISSCOR, Intragenic, Stochastic Synonymous Codon Occurrence Replacement; PCA, principal component analysis.

by analyzing the relative synonymous codon usage and the codon substitution patterns. Wong and coworkers (2010) examined codon usage in the genes of avian and human influenza viruses, including pandemic 2009–2010 H1N1v serotype using the correspondence analysis. They found that the codon usage patterns of seasonal human influenza viruses were distinct among their subtypes, and different from those of avian viruses. Firnberg and Ostermeier (2013) examined the adaptive theory's predictions that for the non-synonymous mutations the average fitness cost of point mutations should be lower than that of two base pairs (bp) to three bp substitutions. This have been done by systematically studying the distribution of fitness effects of 1896 unique single amino acid substitutions in the two genes modified through a combination of computational design and directed evolution. The genes in question coded for inhibitors of HA interaction with target cells. Whitehead and coworkers (2011; 2012) extended these studies, by creating degenerate codon libraries, such that the first two nucleotides in a codon can be any base, but the third nucleotide was limited to G or T, to reduce the frequency of nonsense codons while still allowing all possible amino acids, and then studied their effects on HA binding interactions.

Recently we have proposed an *in silico* method (Radomski & Slonimski, 2009) to tackle the problem of the sequential order of synonymous codons, called ISSCOR (Intragenic, Stochastic Synonymous Codon Occurrence Replacement). In this approach synonymous codons, which occur at different positions of an ORF are replaced randomly by a Monte Carlo routine with their equivalents — the method generates nucleotide sequences of non-original ORFs, which have identical codon usages, and would encode identical amino acid sequences. The ISSCOR method was then used to analyze temporal and spatial aspects of the three sets of orthologous gene sequences isolated from various strains of hemagglutinin of the influenza A virus subtypes: A/H3N2, A/H1N1 (of both the seasonal, and the 2009 pandemic variants), and A/H5N1 (Radomski & Slonimski, 2012) in an alignment-free manner. Coleman and coworkers (2008) has also studied a similar sequential codons' order problem in many viral systems, including influenza. However, they limited the scope to examine only the neighboring codon pair usage — in their studies the amino acid sequence and codon bias of viruses was varied by the codon pair usage of the neighboring codon hexamers. In the experiments of Greenbaum and coworkers (2008) randomized viruses were used, maintaining their amino acid sequence and codon usage, which were then compared to that of the original real virus – the comparisons were also made across influenza strains in different hosts.

The rich collection of the data gathered during the last swine flu pandemic H1N1 permitted a fresh look on the perennial questions of influenza epidemiology. The role of founder effects is important for epidemiological scenarios (Nelson *et al.,* 2009; Lee *et al.,* 2010), assuming that a genetic variability common to a small founder population will then also be found in most descendants. In viral outbreaks such effects can be at play when specific mutations are enriched in samples coming from the same region, and/or the same period. Considering phylogenetic relations it is useful to identify such viral lineage founder events. The global strain sequencing efforts, combined with robust statistics allow novel insights into phylogeny, and especially variability of this highly changeable RNA virus. The variability problem is of interest in view of recent controversy concerning the switching of receptor selection by the hemagglutinin (Imai *et al.,* 2012; Herfst *et al.,* 2012), leading to a possible acquiring of airborne infection transmissibility for mammalian hosts. Based on their combined results, Russell and coworkers (2012) proposed a mathematical model of within-host H5N1 virus evolution to study some aspects influencing increase or decrease in probability of subsequent substitutions leading to aforementioned switch. The authors stressed that more data are needed for assessing calculated evolution rates based on the assumed mutation rates, the problem of high interest for assessing the speed of evolution of HA towards switching receptor selection. We postulated in turn that rates of mutation frequencies in HA commonly accepted are routinely overestimated for at least one order, and proposed an enhanced method of finding evolutionary correlations between multiple strains of the H1N1 2009 pandemic virus (Radomski *et al.,* 2014).

The goal of the current work is to explore in some detail relationships between different orthologous hemagglutinin sequences of various influenza A serotypes, using their ISSCOR descriptors. As these descriptors are relatively easy to calculate, and yet encompass a rich spatial information involving long range interactions together with immediate neighborhoods of constituting residues, it should be possible to correlate stochastic genes' representation based upon the deviates, and their possible 3D epitopic functional interactions — either through theoretical 3D modeling, or perhaps through building associations with appropriate biological data.

The analysis of large and very large collections of nucleotide sequences is a difficult problem, concerning especially reconstruction accuracy and its dependence on sampling rates (Ranala *et al.,* 1998; Ho & Jermiin, 2004; Edwards & Rausher, 2009; Than & Nakleh, 2009; Leache *et al.*, 2011). Two major underlying reasons are**:** the incomplete lineage sorting (ILS), and especially a possible absence in the analyzed sequences set some of key missing ancestors (MA). The problem of ILS received a lot of attention in the field of inferring species trees from gene trees. However, some issues leading to branches entanglement are also common in unraveling topology for a strain tree. In particular, when distances of sequence pairs are estimated within a set, it often happens that a given sequence might be less distant to some other sequence from an entirely different evolutionary pathway, than to its actual immediate ancestor, leading to its wrong branch assignment. For more details and an overview of other issues involved c.f. Radomski and coworkers (2014), where we analyzed a large set of over three thousands of unique HA H1N1 2009–2010 pandemic sequences, using the modified neighbor joining (NJ+; Płoński & Radomski, 2010; 2013) and some related phylogenetics techniques. Already for the resulting phylogenetic graphs their ease of handling, clarity of results, and a general usability were quite limited, and here the analyzed set is about three times as large. Therefore in the current study an application of classic phylogenetic approaches was not even attempted.

MATERIALS AND METHODS

The full length gene sequence of the influenza A hemagglutinin, isolated mostly from avian, human and swine hosts, for serotypes: H1N1 (seasonal), H1N1 (2009–2010 pandemic), H1N2, H2Nx (all neuraminidases), H3N2, H5N1, and H7Nx (all neuraminidases) were obtained on December 9th 2012, from the NCBI influenza resource. From this collection unique sequenc-

Table 1. The distribution of the number of sequences for each serotype per host.

Serotype	Host	Sequences
H1N1s	avian	116
H1N1s	human	1280
H1N1s	swine	678
H1N1v	avian	4
H1N1v	human	3243
H1N1v	swine	197
H1N2	avian	10
H1N2	human	21
H1N2	swine	58
H2Nx	avian	108
H2Nx	human	5
H3N2	avian	98
H3N2	human	2356
H3N2	swine	292
H5N1	avian	539
H5N1	human	107
H5N1	swine	17
H5N1	ferret	2
H7Nx	all hosts	88
All serotypes		9131

es were selected, such that from each subset of identical genes, the ones with the earliest dates of sample isolation were chosen as representatives. Table 1 shows the distribution of the number of sequences for each serotype per host.

Overview of the ISSCOR approach

Previously (Radomski & Slonimski, 2001; 2007), we have described alignment free approaches to the problem of comparison and analysis of complete genomes, and some techniques enabling to cope with the sparseness of the n-gram type (Radomski & Slonimski, 2007; Damashek, 1995) of genomic information representations. The problem of sparse occurrence matrices is not only present, but even more pronounced when dealing with the number of permutations of the possible synonymous codons. Calculating the set of n–grams for such occurrences will lead to a severely sparse vector representations, especially for higher n-grams lengths, and hence to very poor statistics. To alleviate this problem, we proposed (Radomski & Slonimski, 2009) a hybrid approach. Namely, when computing counts of codon-pair patterns — separated by codon sub-sequences of differing length — the actual composition of these spacer sub-sequences will be neglected. When such partial counts are used as a composite set, poor statistics is no longer a hindering obstacle, and the complete information about particular n-gram frequencies profile is preserved, albeit in a distributed and convoluted form.

For every protein coding gene, with its original nucleotide sequence j_0, a set of equivalent nucleotide strings ($j_1, j_2, j_3, \ldots, j_N$) is created by a Monte Carlo approach. These artificial sequences have the following properties:

- they are all of the same nucleotide lengths as the j_0;
- they have exactly the same amino acid sequence as the j_0 (i.e., the proteins translated from the $j_1, j_2, j_3, \ldots, j_N$ are identical to j_0);
- they have in the vast majority of cases a synonymous codon order different from the original sequence j_0.

Therefore, the ISSCOR method allows comparing the original codon sequence with an ensemble of different synonymous sequences — yet all of them coding for the same sequence of amino acids.

The Computational Procedure

Full description of the method is given in (Radomski & Slonimski, 2009), but mathematical steps are briefly outlined here for convenience. First, the codon usage frequencies are determined, and on that basis the probabilities of replacement are calculated, separately for each codon-degeneracy equivalence group E:

$$P^k = \frac{u_k}{\sum_{d=1}^{E} u_d} \qquad (1)$$

where:
the P^k — probability that any other codon from the same degeneracy equivalence group E — will be randomly replaced by the codon k; and u_k is the synonymous codon k triplet frequency for a given amino acid in a whole gene. Therefore, obviously for any given degeneracy equivalence group E, the sum of such probabilities will always be equal to 1. Then, successively for each codon in a gene the procedure of it's synonymous random replacement is performed based on probabilities according to the equation (**1**). Finally, the resulting shuffled sequences are determined, and compared to the original sequence of the gene.

For each protein coding sequence we need to determine a complete matrix of all codon-pair patterns. Obviously, in a protein coding sequence, there are at most 3904 (61×61 + 61×3) unique codon-pair patterns. In order to calculate observed values of a particular codon-pair pattern (c_k, c_l) for a given sequence V and the all codon-spacer lengths, first we need to construct a series of matrices O^λ (*occurrence matrices*). Each element of every matrix O^λ contains the counted sum of all specific codon-pair patterns (c_k, c_l), separated by a string of other codons present in this sequence, where the λ denotes the number of other codons separating the given codon-pair pattern (c_k, c_l). Using a sliding window of the length $3*(\lambda+2)$ nucleotides, and starting at the position m, we would scan the whole sequence V, calculating elements of the matrix by the formula:

$$O^\lambda(c_k, c_l, p) = \sum_{m=1}^{M-\lambda-1} f(c_k, c_l, \lambda, m, p) \qquad (2)$$

where M is the sequence's length, and

$$f(c_k, c_l, \lambda, m, p) = \begin{cases} \textbf{1}, \text{ if } V(m) = c_k \text{ and } V(m+\lambda+1) = c_l, \text{ and the codon-pair pattern } (\boldsymbol{c_k}, \boldsymbol{c_l}) \text{ matches the pattern of the particular comparison } \boldsymbol{p} \\ \textbf{0}, \text{ otherwise.} \end{cases}$$

Comparisons involve matches between the predefined codon-pair patterns, of the first codon c_k always taken together with the second codon c_l. That is, a particular positional comparison p involves only one nucleotide from the first codon c_k, and one nucleotide from the second

codon c_i, ignoring all four remaining nucleotides, which corresponds to a pattern (for convenience we name each such pattern a *hexon*). Thus, there are e.g., nine patterns containing the adenine (**A**) at any position in a first codon, together with the cytosine (**C**) at any position in a second codon, etc. Obviously, when $\lambda = 0$, one has an adjacent codon-pair pattern (hexanucleotide), for $\lambda = 1$ it is a nonanucleotide, and so on. Note, that since these are ordered counts, each starting at the sequence's 5'–terminus, the matrices O_i^{λ} are not symmetrical, that is the count of the pair (c_k, c_l) is different from the count of the pair (c_l, c_k).

To make the results independent of a particular sequence size (or a set of sequences, as described already on the example of the complete genome of *Helicobacter pylori* by Radomski & Slonimski, 2009), we need to calculate how much the number of actually observed *hexons* in the original sequence, differs from the mean number of the corresponding *hexons*, observed after performing N number of random ISSCOR permutations, divided by the standard deviation observed in the corresponding shuffled sequences:

$$D_{xAx\_\lambda\_Txx} = \frac{O_{occurences} - \frac{\sum_{n}^{N} R^{n}_{shuffled}}{N}}{STD_{shuffled}} \qquad (3)$$

where:
$D_{xAx\_\lambda\_Txx}$ is expressed in STD units (termed in mathematic as a *deviate*) for, e.g., the pattern *xAx_λ_Txx*, that is for the all codon combinations comprising the nucleotide *A* at the second position in the first codon, and the nucleotide *T* at the first position of the second codon — the border codons being separated by the number λ of other codons;
$O_{occurences}$ are the numbers of the actually observed occurrences for any given hexon in the unperturbed sequence;
$R^{n}_{shuffled}$ are the numbers of occurrences for any given hexon pattern counted after codons of the whole sequence have been shuffled randomly (as described above), thus the $R^{n}_{shuffled}/N$ is a mean number of such occurrences after the N such random shuffles;
$STD_{shuffled}$ is a standard deviation for occurrence of a given hexon pattern, after N random shufflings of the whole sequence; we have determined previously that 500 shuffles are sufficient to obtain systematic and highly repetitive results (Radomski & Slonimski, 2009).

Phylogenetic analyses were performed using the classic Neighbor Joining (Tamura *et al.,* 2004; Waterhouse *et al.*, 2009), and the modified NJ algorithm: QPF (Płoński & Radomski, 2010). The results from the II-iteration trees after multiple alignment, available through the MUSCLE package of Edgar (2004), were checked for consistency with distance-based methods. For tree manipulations (in the Newick format) and their visualization, the Dendroscope package of Huson and coworkers (2007) was used.

RESULTS AND DISCUSSION

The ISSCOR deviates (equation 3) for all the 9131 hemagglutinin sequences collected were calculated as described earlier, using. codon spacer values of 0 to 16 for λ, and creating the matrix MA of 9131 rows by 2448 columns. The results of principal component analysis (PCA) for the matrix MA showing 9131 data points (marked in light gray), are presented in Figs. 1 to 3, showing maps of PC-1 values (45.4% of a total variance explained) — plotted on the abscissas, and the PC-2 values (further 18.3% of a total variance) — plotted on the ordinate axes respectively. The intent is that the 9131 points present in all the maps will provide a common frame of reference. On such a background the points corresponding to the different serotypic subsets of genes are color-coded, and grouped by the host (avian, human, swine, or few cases of ferrets).

Chronology of evolution

The presumed trend in a chain of infectivity from avian, through porcine, to human hosts can be well observed e.g. in the case of H3N2 serotype on the right hand column panels in Figs. 2A and 2B. Although it is usually considered that the avian hosts viruses form a primary reservoir of infections, and the propagation follows from avian, through mammals (mostly porcine) to human hosts, in case of the 2009 pandemic H1N1 strains isolated from avian hosts, there was most probably a reversal of influenza infectivity chain. Such a conclusion, stemming from the analyses of the maps here, was subsequently confirmed by the search in original literature. There are only four such isolates present in the NCBI database (accessions: HM370960, HM370967, HM370975, HM450134), isolated in Canada from turkeys (Berhane *et al.,* 2010) between Oct. and Dec. 2009. These sequences group on the map together with the cluster of the isolates from human and swine hosts (the left column on Fig. 1).

In contrast to the H3N2 strains, which all form together one large, elongated cluster, with a clearly visible evolutionary trend (right side of Figs. 2A and 2B; c.f. also Radomski & Slonimski, 2012), the behavior of all other isolates (H1N1, H5N1, H1N2, H2Nx, and H7Nx) is much more complex, as they intermingle, forming a network of possible evolutionary paths. It is possible to roughly trace an early chronology of e.g. human and porcine of the seasonal H1N1 isolates (Fig. 1) from the earliest complete gene sequence of 1918, and then successively from 1930', 40', 50', and so on. All the 9131 points do form an approximately triangular shape on each map, with the most recent 2009 H1N1 pandemic strains occupying the leftmost bottom apex, the most recent human isolates of the H3N2 in the rightmost bottom one, and the topmost position occupied by the human seasonal H1N1 strains of the I-st decade of XXI c. (thus the H1N1 strains are forming the left elongated, crescent-shaped edge; however, with some avian and porcine isolates already stretching towards regions occupied by the H3N2 ones).

Assuming that a hypothetical, primeval ancient influenza hemagglutinin might have occupied a position somewhere in the middle of this triangle, we can follow the chronological spreading of more recent strains towards more and more remote locations (c.f. Fig. 1). Therefore, it seems reasonable to assume that HA orthologs occupying positions in apexes of this triangular distribution may mark the extent of this gene changeability in influenza A-type viruses, although of course further genetic shif might expand these boundaries even more. The HA sequences of the three most extended positions on the map are all obtained from isolates of human hosts**:** the topmost — A/Hamburg/1/2005 (H1N1 seasonal; FJ231765), bottom-left — A/Sydney/DD3_17/2010 (H1N1 pandemic, CY092550); and the bottom-right — A/Thailand/Siriraj-06/2002 (H3N2, JN617982), These observations are consistent with the hypothesis that all influenza viruses originated from their ancient stock har-

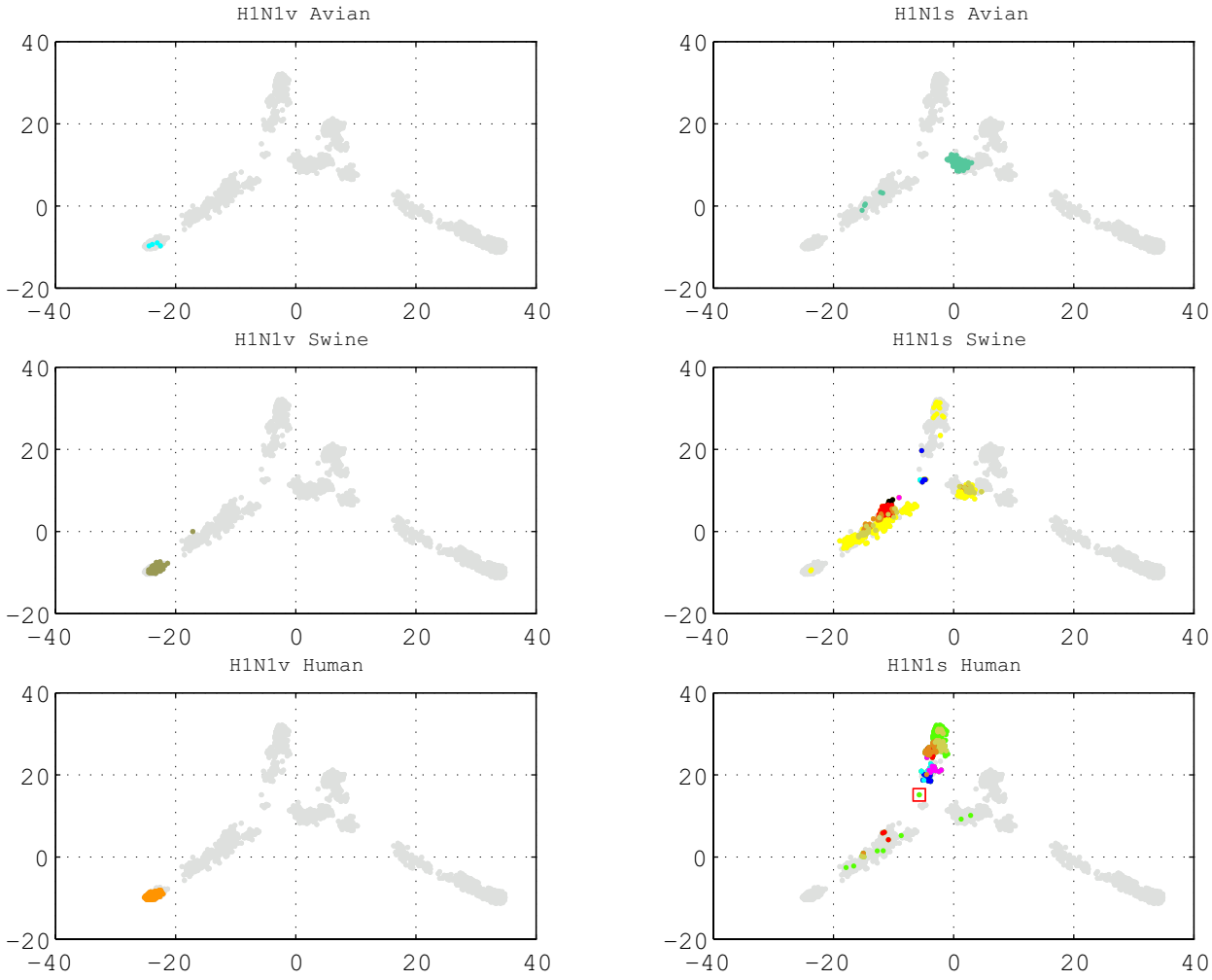

Figure 1. PC-1 *vs*. PC-2 scatter-plot of principal component analysis of the ISSCOR descriptors for the H1N1 seasonal and H1N1 pandemic serotypes
Positions of all the 9131 full-length hemagglutinin sequences analyzed (light gray points) superposed with the sequences corresponding to the 2009–2010 pandemic (H1N1v, left column), and the seasonal (H1N1s, right column) H1N1 serotypes, are color-coded according to their infected host. The early XX century sequences isolated from the swine (in the middle-right panel) and the human (in the bottom-right panel) are color coded as follows: a case of "Spanish flu" (strain A/South_Carolina/1/1918) — red square; sequences from 1930′ – blue; from 1940′ — cyan; from 1950′ – magenta; from 1960′ — black; from 1970′ — red; from 1980′ — brown; from 1990′ — khaki; and from 2000′ and 2010′ — light green (human) and yellow (swine) points.

Table 2. The distance matrix, showing the respective numbers of nucleotide differences between some representative strains.

Accession	CY008988	CY125862	CY026283	CY020381	CY087800	CY087792
Serotype	H1N1 (human)	H1N1 (human)	H1N1 (swine)	H2N2 (human)	H2N2 (human)	H2N2 (human)
Strain	A/Denver/1957	A/Kw/1/1957 (China)	A/swine/Wisconsin/1/1957	A/Albany/26/1957	A/Japan/305-MA12/1957	A/Singapore/1-MA12E/1957
A/Denver/1957	0	15	1012	1218	1220	1218
A/Kw/1/1957	15	0	1012	1219	1221	1219
A/swine/Wisconsin/1/1957	1012	1012	0	1218	1217	1215
A/Albany/26/1957	1218	1219	1218	0	6	3
A/Japan/305-MA12/1957	1220	1221	1217	6	0	7
A/Singapore/1-MA12E/1957	1218	1219	1215	3	7	0

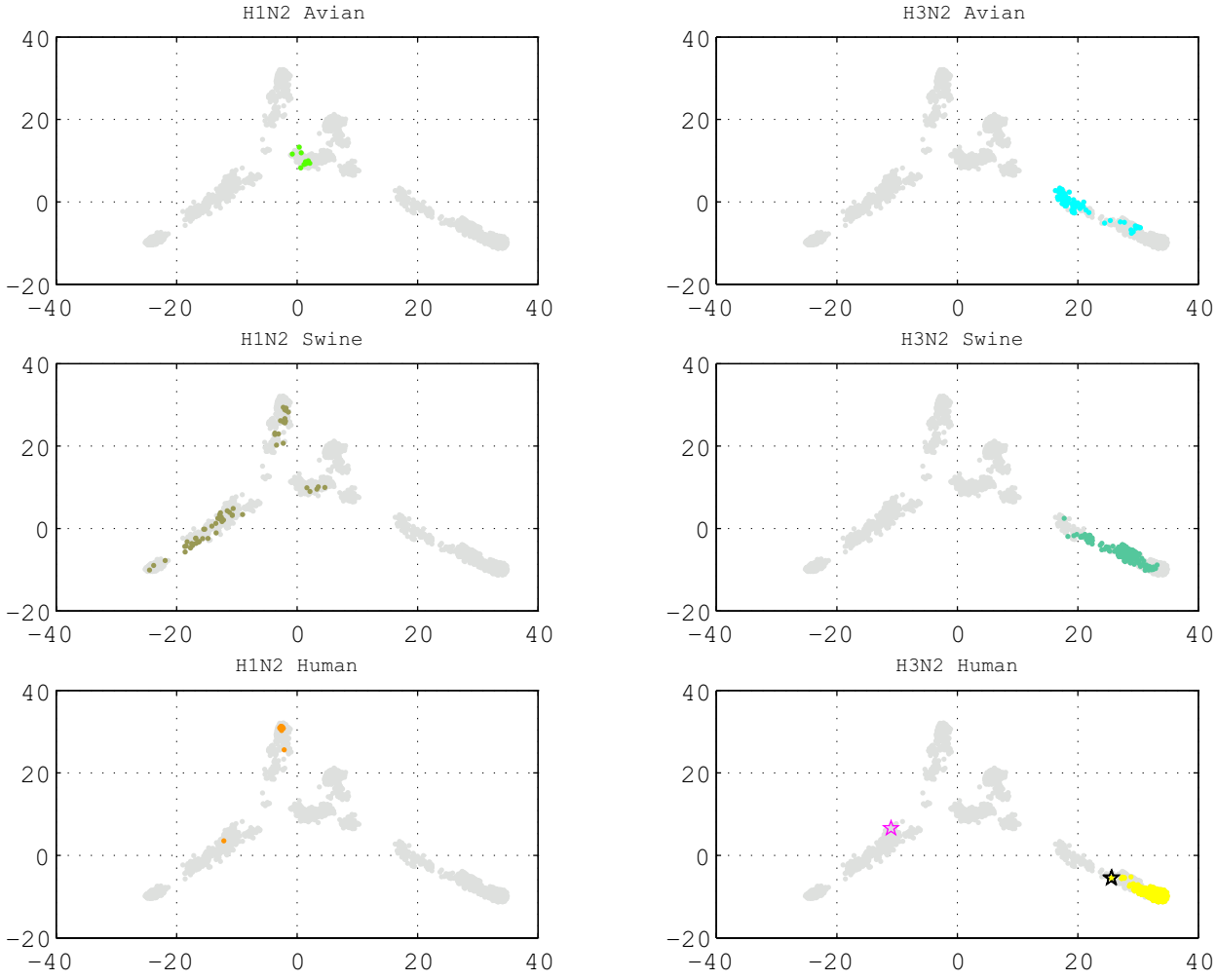

Figure 2 (Panel A). PC-1 *vs.* PC-2 scatter-plot of principal component analysis of the ISSCOR descriptors for the H1N2 and H3N2 serotypes
Positions of all the 9131 full-length hemagglutinin sequences analyzed (light gray points) superposed with the sequences with the sequences corresponding to the H1N2 (left column), and H3N2 (right column) serotypes, are color-coded according to their infected host. Additionally: on the bottom-right panel — the reference Hong Kong 1968 pandemic strains are marked: A/Hong_Kong/1-10-MA21-1/1968 (CY080523, black star), A/Hong_Kong/1-4-MA21-1/1968 (CY080515, black star), as well as the A/swine/Wisconsin/1/1968 (EU139825, of H1N1 subtype; magenta star).

bored in wild, migratory aquatic birds. Accordingly the avian host isolates are, on average, closest to the center of the triangular spread of HA genes observed here, followed by porcine isolates, and only then isolates from humans. In agreement with an assumption that considers swine acting often as an intermediate host.

Putative origins of the XX c. pandemics

The question arises then whether it might be possible to trace back an origin of genetic shifts in HA leading to known major pandemic outbreaks. There is a complete dearth of data of sequences prior to the H1N1 1918 strain. Also, for the next major pandemic emergence in 1957 the H2N2 viruses that caused the Asian flu, but disappeared from human population a decade later. There are only six complete HA H2N2 genes present in this set, which is not sufficient to draw valid conclusions. The Table 2 contains the distance matrix, showing the respective numbers of nucleotide differences between these strains (c.f. also Fig. 3A and 3B, top-right panels).

Already in 1993, the most probable avian origin of the Asian flu pandemic have been determined on the basis of antigenicity of H2 HAs from representative human and avian viruses, as well as of their evolutionary characteristics in respective hosts (Schafer *et al.,* 1993).

Supposedly, the human H2 HAs, that circulated in the 1957–1968 period, formed a separate phylogenetic lineage, most closely related to the Eurasian avian H2 HAs, while the antigenically conserved counterparts of the human Asian pandemic strains of 1957 might still continue to circulate in the avian reservoir, continuously coming into a close proximity with susceptible human populations. There was also an increased prevalence of H2 influenza viruses among wild ducks in 1988 in North America (Schafer *et al.,* 1993), preceding the appearance of H2N2 viruses in domestic fowl. As the prevalence of avian H2N2 influenza viruses increased on turkey farms and in live bird markets in New York City and elsewhere, greater numbers of these viruses have come into direct contact with susceptible humans. Unfortunately, the earliest avian host full HA sequence in our 9131 set

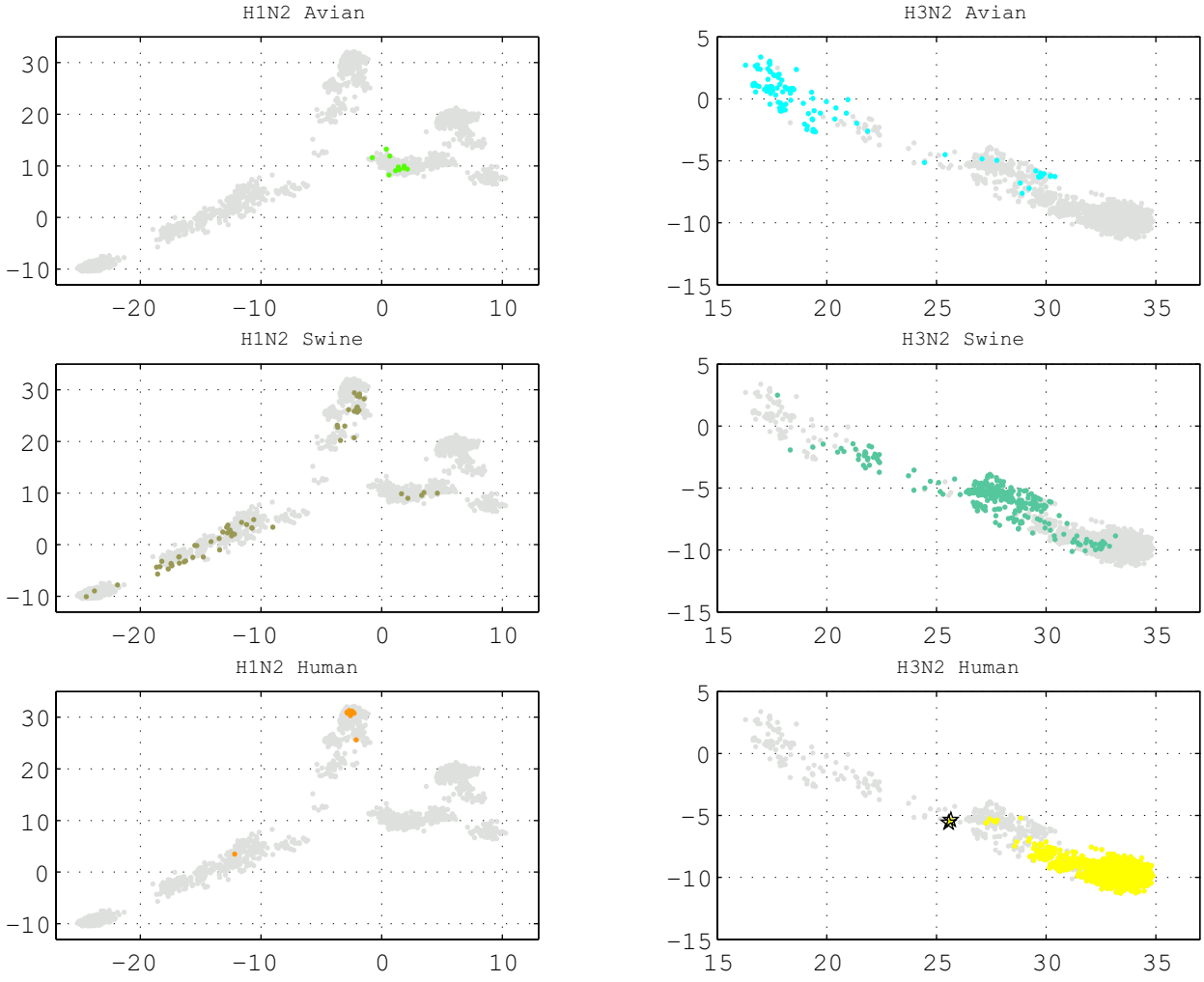

Figure 2 (Panel B). PC-1 *vs.* PC-2 scatter-plot of principal component analysis of the ISSCOR descriptors for the H1N2 and H3N2 serotypes
Positions of all the 9131 full-length hemagglutinin sequences analyzed (light gray points) superposed with the sequences corresponding to the H1N2 (left column), and H3N2 (right column) serotypes displayed here are the same as on Fig. 2A, but now showing only the PC-1 and PC-2 regions where the respective H3N2 orthologs were present. Additionally: on the bottom-right panel — two reference Hong Kong 1968 pandemic strains are marked as black stars: A/Hong_Kong/1-10-MA21-1/1968 (CY080523), A/Hong_Kong/1-4-MA21-1/1968 (CY080515).

was isolated only in the 1969, so we can't infer the 1957 epidemic origin from the current ISSCOR map.

Then in July 1968 the next pandemic's virus H3N2 was first isolated in Hong Kong (Cockburn *et al.,* 1969). Again, there are only three strains isolated in 1968 among 9131 HAs: A/Hong_Kong/1-10-MA21-1/1968 (CY080523), A/Hong_Kong/1-4-MA21-1/1968 (CY080515), and A/swine/Wisconsin/1/1968 (EU139825, of H1N1 subtype). The first two are closely related (six mutations distant from each other), while the third one is distant from both by about 1200 nucleotide differences, and it does not seem to be a pandemic precursor. Scholtissek and coworkers (1978) concluded that the H3N2 subtype presumably derived from a H2N2, by retaining seven segments of the H2N2, while the gene coding for the HA was recombined from Ukrainian duck or another highly related avian strain (not present on the ISSCOR map here).

A detailed ISSCOR-PCA analysis to illustrate more clearly the method's possibilities to e.g. pinpoint distribution of the putative mutations necessary to be present in order to the H5N1 undergo the transitions described by Herfst and coworkers (2012) from the avian transmissible to the mammals airborne-transmissible can be found it the Supplementary Materials (at www.actabp.pl, the "Laboratory-induced transition to the droplet-transmissible infectivity" Section).

The swine-like H1N1 pandemic virus, 2009–2010

In contrast to the previous outbreaks, during the 2009 H1N1 pandemic "swine flu" an abundant collection of data was gathered all across the globe.

The NJ+ phylogram of the 2009 pandemic H1N1 HA all early 178 sequences (isolated early in the 2009, beforeApril 30th) is shown in The Fig. S1 (Supplementary Materials at www.actabp.pl). The tree was constructed using also the 380 putative precursor sequences of other serotypes (all collected during the same period). The two most probable swine precursors: the H1N1 A/swine/Missouri/46519-5/2009 (HQ378741) and the H1N2 A/swine/Hong_Kong/NS252/ 2009 (CY085998), form a small sub-clade, adjacent to the newly emerged pan-

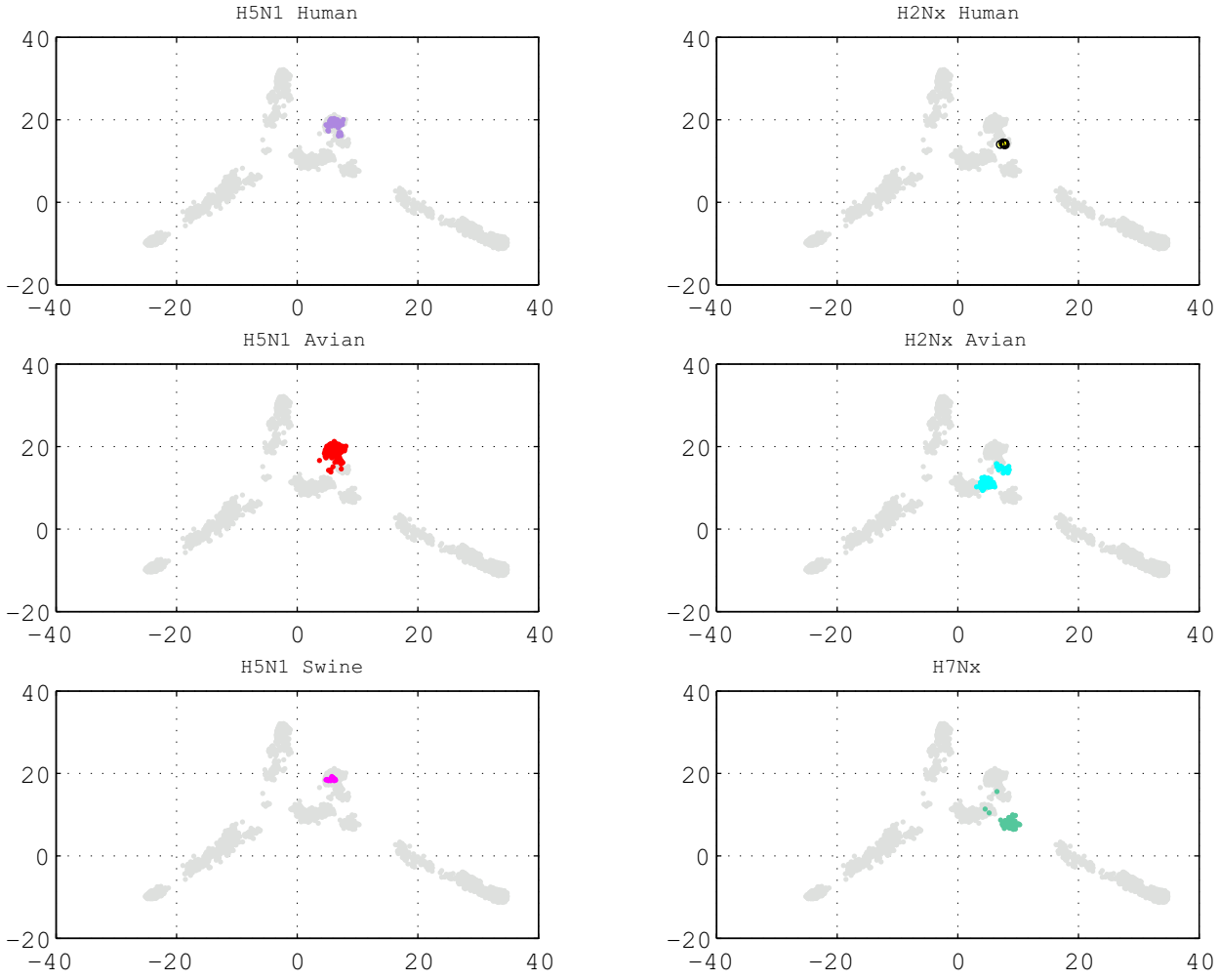

Figure 3 (Panel A). PC-1 *vs.* PC-2 scatter-plot of principal component analysis of the ISSCOR descriptors for the H5N1, H2Nx and H7Nx serotypes
Positions of all the 9131 full-length hemagglutinin sequences analyzed (light gray points) superposed with the sequences corresponding to the H5N1 (left column), H2Nx (right column, top and middle), and H7Nx (right column, bottom) serotypes.

demic strains (which form together one tight cluster) but that is distant from all the other possible ancestors. The CY085998 is a triple reassortant swine strain, present in a large study of phylogenetic evolution relationships of putative precursors to the 2009 H1N1 pandemic. This strain was examined in great details in Vijaykrishna D *et al.,* (2011), and although it wasn't indicated as a most probable human pandemic precursor by the authors, its position on their NJ tree (c.f. Fig. S2a in Supplementary Materials of Vijaykrishna *et al.,* 2011) marks it as a very likely candidate.

As the incubation and infectivity period of the virus lasts about one week, it is clear that the precursors ought to be extant at a time of possible genetic shift event to occur, however, we did repeat the analysis including also as an additional candidates all the 569 strains isolated during 2008. Partial results are shown on Fig. S2 (Supplementary Materials at www.actabp.pl), besides two sequences described, there were also three other porcine H1N1 strains, preceding the HQ378741, in the same small sub-clade: the A/swine/North Carolina/3793/2008 (JQ624667), the A/swine/Illinois/02064/2008 (CY099095), and the A/swine/Ohio/02026/2008 (CY099159), confirming validity of the former analysis. The relationships between these five putative precursor sequences on the ISSCOR map are shown on Fig. S3 (Supplementary Materials at www.actabp.pl).

The pandemic 2009 H1N1 cluster contains also five other porcine-host HA genes, but they are all more recent than May 2009: the H1N2 (circles) — A/swine/Italy/116114/ (CY067662), A/swine/Minnesota/A01076209/2010 (JQ906868), and A/swine/Nebraska/A01203626/2012 (JX444788); and the H1N1 (crosses) — A/swine/Illinois/ A01076179/2009 (JX042553, isolated Dec. 6th, 2009), and A/swine/Shepparton/6/2009 (JQ273542, isolated August 17th, 2009). In line with our observations, Garten *et al.,* (2009) have found that molecular markers predictive of adaptation to humans were not present in the early (as of May 2009) pandemic H1N1 viruses, and that antigenically the viruses were homogeneous — similar to North American swine H1N1 viruses, but distinct from seasonal human H1N1 isolates.

Similarly, Smith and coworkres (2009) concluded that the initial transmission to humans must have occurred several months before recognition of the outbreak. Moreover, the unsampled history prior to pandemic means that the nature and location of the genetically closest swine viruses revealed little about the immediate origin of the epidemic.

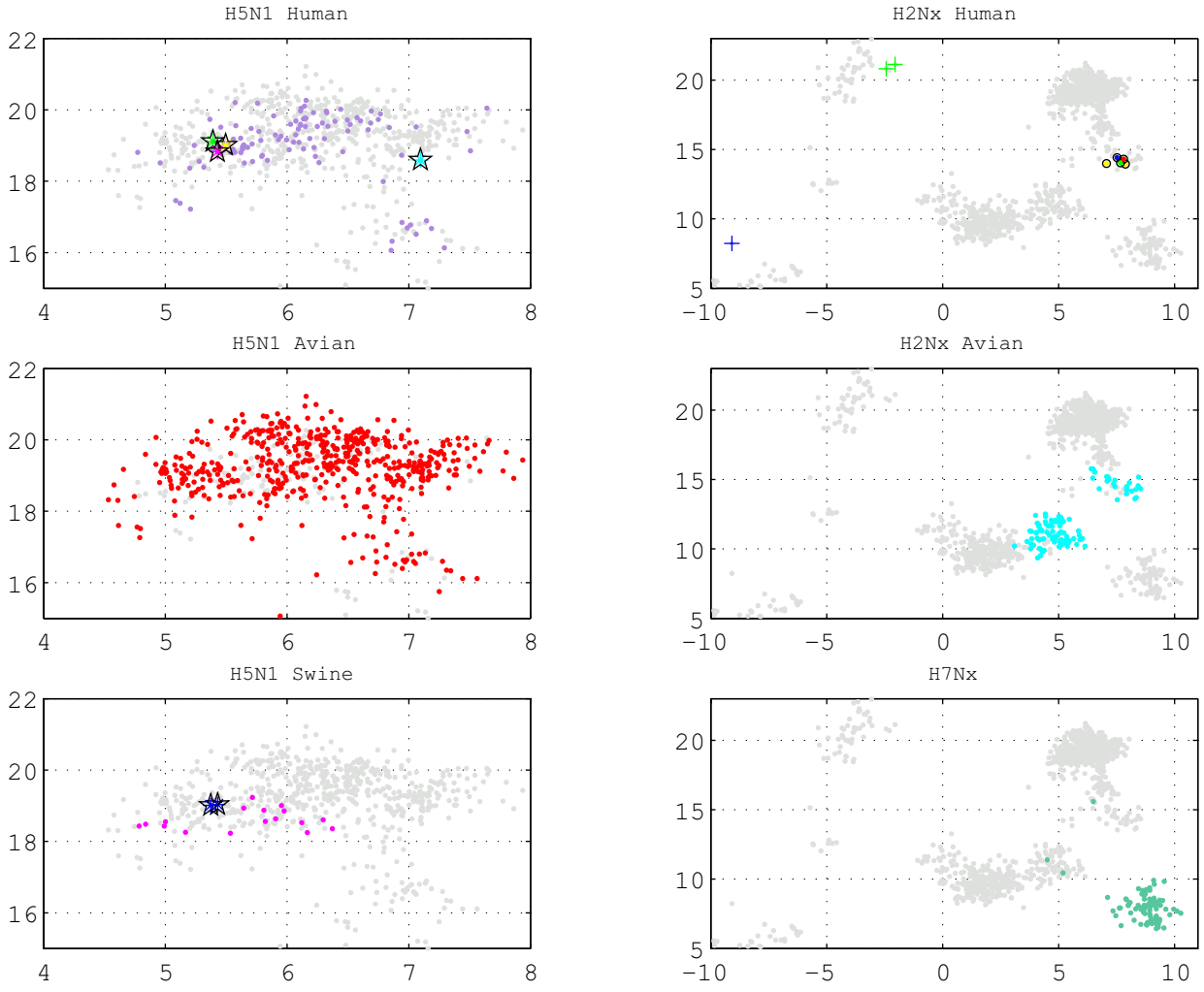

Figure 3 (Panel B). PC-1 *vs.* PC-2 scatter-plot of principal component analysis of the ISSCOR descriptors for the H5N1, H2Nx and H7Nx serotypes
Positions of all the 9131 full-length hemagglutinin sequences analyzed (light gray points) superposed with the sequences corresponding to the H5N1 (left column), H2Nx (right column, top and middle), and H7Nx (right column, bottom) serotypes — same as on Fig. 3A, but showing only the PC-1 and PC-2 regions where the respective H5N1 and H2N2 orthologs were present. Additionally: on the top-left panel — the reference wild type strains of the A/Indonesia/5/2005 isolated from human hosts are marked individually (accessions: CY116646 — yellow, the Fouchier's sequence from Herfst et al., (2012); EU146622 — green; GQ149235 — magenta; and AY651334 — cyan, the /VietNam/1203/2004). The two blue stars in the bottom-left panel belong to the Fouchier's ferret sequences (CY116654 of their ferret #1, and CY116662 for ferrets #2 to #7); and on the top-right panel — the reference sequences of the 1957 Asian flu H2N2 (c.f. also Table II) are marked: CY087792 A/Singapore/1-MA12E/1957 — blue; CY087800 A/Japan/305-MA12/1957 — green; CY020381 A/Albany/26/1957 — red; as well as 1957 sequences of H1N1 serotype: CY008988 A/Denver/1957 and CY125862 A/Kw/1/1957 — green crosses (human), and CY026283 A/swine/Wisconsin/1/1957 (porcine).

Surprisingly, if we compare an over 40 years long dominance of the H3N2 serotype, the new H1N1 variant displayed already at only two years since the 2009 pandemic outbreak rather low abundance. Of the 190 full length HAs isolated from porcine hosts during eleven months of 2012 till December 9th there were 115 of the H1N1 seasonal serotype, 2 of H1N2, 34 of the H3N2; from avian hosts: 10 of the H5N1, and 1 of the H7N3; from human hosts: 1 of the H7N3, 3 of the H5N1, 14 of the H3N2; however, there were only 10 of the recently dominant H1N1 2009–2010 pandemic type strains.

CONCLUSIONS

The ISSCOR-PCA method enables fast and efficient visualization of evolutionary relations present in a very large, complex set of homologous sequences. It is otherwise not an easy and rather tedious task, when applying other phylogenetic analysis algorithms available to thousands of sequences. Our approach significantly simplifies the effort, by producing two-dimensional projections from the multidimensional hyperspace of descriptors characterizing each of individual strains, allowing a clear understanding of the genetic diversity inside such large set of homologs.

Based on the sequences 9131 set, we have examined the odds of putatively tracking an origins of all flu pandemics in XX century, however, their mostly unsampled history and especially a severe paucity of zoonotic data for all major genetic shifts prior to the H1N1 pandemic of 2009–2010, deemed the task not possible. In contrast, for the 2009 "swine flu" H1N1 pandemic, an abundant collection of sequential data was gathered from human hosts, but again not so many from the porcine or avian

ones. The ISSCOR maps confirmed the close affinity of the earliest HAs of human isolates to their tentative precursors from swine, and yet even for this well documented epidemic the nature and location of the genetically closest swine viruses reveal little about the immediate origin of the infection. Clearly, much higher ratio of porcine and avian isolates needs to be routinely monitored in future, to feasibly pinpoint inter-species acts of transmission, even if only in *ex post* descriptive a manner.

Tracing the chronology of individual strains isolation times on the serotype-specific maps revealed that oldest strains occupy mostly positions in the middle of the roughly triangular shape distribution (Fig. 1), whereas newer strains spread gradually towards apexes of that triangle. Importantly, our analysis shows that the most distant sequences of hemagglutinin were all isolated from human hosts: A/Hamburg/1/2005 (H1N1 seasonal), A/Sydney/DD3_17/2010 (H1N1 pandemic 2009-2010); and A/Thailand/Siriraj-06/2002 (H3N2). Unexpectedly the ISSCOR analysis showed that the hemagglutinin variability is largest in case of strains invading humans, and seems to be less pronounced in case of strains detected in birds. However, as sampling is much skewed towards human strains (c.f. Table 1), the statistics of the collected set do not allow for any far–reaching hypotheses concerning species' specific virus–host interactions. Nevertheless, taking into account the sheer volume of the data analysed we propose that the edges on the ISSCOR maps of this assembly might delimit the extent of genetic diversification of the influenza virus hemagglutinin. This bears on the immunological variability of the HA gene, allowing for a broader look on influenza epidemiology.

In particular, large amounts of systematically collected data available now through the NGS experiments (in particular from an *ab initio* type of laboratory settings) are calling for a rapid yet efficient examination of their results. For example, in an interesting study Renzette and coworkres (2012) examined *ab initio* passaging of the A/Brisbane/59/2007 (H1N1), and the A/Brisbane/10/2007 (H3N2) in MCDK cell cultures, followed by a deep sequencing study, and demonstrated that some surprises might await there, as they have shown rather unexpected *increase* in both serotypes' viral diversity, occurring at the same time after the eights passage, albeit in two separate cell lines. Later on, a related work of Foll and coworkres (2014) successfully solved a challenge of distinguishing genetic drift from selection by a time-sampled evaluation of the whole-genome trajectories of influenza A H1N1 evolution in the presence and absence of oseltamivir, using NGS sequence-rich collections. As such deep sequencing *ab initio* experiments increasingly often produce very large amounts of reliable evolutionary data, it would be of high interest to study them in much more detail by applying the ISSCOR-PCA method presented here; optimally also e.g. in conjunction with ancestry tracking based on disentangled phylogenetic graphs (Radomski *et al.,* 2014) — in our opinion it would form a very promising venue for a near future, albeit also a truly challenging test bed for this approach.

Acknowledgements

We are grateful to Pat Churchland for looking over the English.

This work was partially supported (for JPR and PPS) by the EU project SSPE-CT-2006-44405, and the grant (for JPR) from the National Science Center under the decision DEC-2013/09/B/NZ2/00121. Additional partial funding was generously provided by the WND-POIG.01.01.02-00-007/08 grant from the European Regional Development Fund.

Conflicts of interests

Authors declare no conflict of interests.

REFERENCES

Adzubei AA, Adzubei IA, Krashennikov IA, Neidle S (1996) Non-random usage of 'degenerate' codons related to protein three-dimensional structure. *FEBS Lett* **399**: 78–82.

Ahn I, Son HS (2012) Evolutionary analysis of human-origin influenza A virus (H3N2) genes associated with the codon usage patterns since 1993. *Virus Genes* **44**: 198–206, doi:10.1007/s11262-011-0687-4.

Antezana MA, Kreitman M (1999) The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences, *J Mol Evol* **49**: 36–43.

Berg OG, Kurland GC (1997) Growth rate-optimized tRNA abundance and codon usage. *J Mol Biol* **270**: 544–550.

Berhane Y, Ojkic D, Neufeld J, Leith M, Hisanaga T, Kehler H, Pasick J (2010) Molecular characterization of pandemic H1N1 influenza viruses isolated from turkeys and pathogenicity of a human pH1N1 isolate in turkeys. *Avian Dis* **54:** 1275–1285.

Buchan JR, Aucott LS, Stansfield I (2006) tRNA properties help shape codon pair preferences in open reading frames. *Nucl Acids Res* **34**: 1015–1027.

Carbone A, Kepes AF, Zinovyev A (2005) Codon bias signatures, organization of microorganisms in codon space and lifestyle. *Mol Biol Evol* **22**: 547–561.

Cockburn WC, Delon PJ, FerreiraW (1969) Origin and progress of the 1968–69 Hong Kong influenza epidemic. *Bull World Health Org* **41**: 345–348.

Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S (2008) Virus attenuation by genome-scale changes in codon pair bias. *Science* **32**: 1784–1787.

Collins DW (1993) Relationship between G+C in silent sites of codons and amino acids composition of proteins. *J Mol Evol* **36**: 201–213.

Damashek M (1995) Gauging similarity with n-grams: language-independent categorization of text. *Science* **267**: 843–348.

Dries M, Savva R, Wernish L (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucl Acids Res* **32**: 5036–5044.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatic* **5**: 113; doi:10.1186/1471-2105-5-113.

Edwards SV, Rausher M (2009) Is a new and general theory of molecular systematics emerging? *Evolution* **63**: 1–19.

Eyre-Walker A, Bulmer M (1993) Reduced synonymous substitution rate at the start of enterobacterial genes; *Nucleic Acids Res* **21**: 4599–4603.

Firnberg E, Ostermeier M (2013) The genetic code constrains yet facilitates Darwinian evolution. *Nucl Acids Res* **41**: 7420–7428; doi:10.1093/nar/gkt536.

Foll M, Poh Y-P, Renzette N, Ferrer-Admetlla A, Bank C *et al.* (2014) Influenza virus drug resistance: a time-sampled population genetics perspective. *PLoS Genet* **10**: e1004185; doi:10.1371/journal.pgen.1004185.

Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, Donis R (2009) Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **325**: 197–201; doi: 10.1126/science.1176225.

Greenbaum BD, Levine AJ, Bhanot G, Rabadan R (2008) Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PloS Pathogens* **4**: e1000079.

Herfst S, Schrauwen EJ, Linster M, Chutinimitkul S, de Wit E, Munster VJ, Fouchier RA (2012) Airborne transmission of influenza A/H5N1 virus between ferrets. *Science* **336**: 1534–1541.

Ho SYW, Jermiin LS (2004) Tracing the decay of the historical signal in biological sequence data. *Syst Biol* **53**: 623–637.

Huson D, Richter DC, Rausch Ch, Dezulian T, Franz TM, Rupp R (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**: 460; doi:10.1186/1471-2105-8-460.

Ikemura T (1985) Codon usage and tRNA content of unicellular and multicellular organisms. *Mol Biol Evol* **2**: 13–34.

Imai M, Watanabe T, Hatta M, Das SC, Ozawa M, Shinya K, Kawaoka Y (2012) Experimental adaptation of an influenza H5HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature* **486**: 420–430; doi:10.1038/nature10831.

Kimura M (1962) On the probability of fixation of mutant genes in populations. *Genetics* **47**: 713–719.

Knight RD, Freeland SJ, Landweber LF (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol* **2**(4): research0010.1-0010.13, doi:10.1186/gb-2001-2-4-research0010.

Leache AD, Rannala B (2011) The accuracy of species tree estimation under simulation — a comparison of methods. *Syst Biol* **60**: 126–137.

Lee RT, Santos CL, de Paiva TM, Cui L, Sirota FL, Eisenhaber F, Maurer-Stroh S (2010) All that glitters is not gold — founder effects complicate associations of flu mutations to disease severity. *Virol J* **7**: 297; doi:10.1186/1743-422X-7-297.

McHardy CA, Adams B (2009) The role of genomics in tracking the evolution of influenza A virus. *PLoS Pathog* **5**: e1000566; doi:10.1371/journal.ppat. 1000566.

Nelson M, Spiro D, Wentworth D, Fan J, Beck E, George KS, Henrickson K (2009) The early diversification of influenza A/H1N1pdm. *PLoS Currents Influenza* **Nov 5**; doi: 10.1371/currents.RRN1126.

Novozhilov AS, Wolf YI, Koonin EV (2007) Evolution of genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol Direct* http://www.biology-direct.com/content/2/1/24.

Płoński P, Radomski JP (2010) Quick Path Finding — quick algorithmic solution for unambiguous labeling of phylogenetic tree nodes. *Comput Biol Chem* **34**: 300–307; doi:10.1016/j.compbiolchem.2010.10.002

Radomski JP, Slonimski PP (2001) Genomic style of proteins: concepts, methods and analysis of ribosomal proteins from 16 microbial, species. *FEMS Microbiol Rev* **25**: 425–435.

Radomski JP, Slonimski PP (2007) Primary sequences of proteins from complete genomes display a singular periodicity: Alignment-free N-gram analysis; *CR Biol* **330**: 33–48.

Radomski JP, Slonimski PP (2009) ISSCOR: Intragenic, Stochastic Synonymous Codon Occurrence Replacement — a new method for an alignment-free genome sequence analysis, *CR Biologies* **332**: 336–350.

Radomski JP, Slonimski PP (2012) Alignment free characterization of the influenza A hemagglutinin genes by the ISSCOR method. *CR Biologies* **335**: 180–193.

Radomski JP, Płoński P, Zagórski-Ostoja W (2014) The hemagglutinin mutation E391K of pandemic 2009 influenza revisited. *Mol Phyl Evol* **70**: 29–36.

Rannala B, Huelsenbeck JP, Yang Z, Nielsen R (1998) Taxon sampling and the accuracy of large phylogenies. *Syst Biol* **47**: 702–710.

Renzette N, Liu P, Caffrey D, Zeldovich K, Palmiotti C, Spain J, Camolli J, Borenstein J, Schiffer C, Wang J, Finberg R, Kowalik T (2012) *In vitro* culture of influenza A virus alters the evolutionary trajectories of viral populations. *15th International Conference on Infectious Diseases,* Bangkok, June 13–16, 2012.

Rodnina MV, Wintermeyer W (2001) Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Ann Rev Biochem* **70**: 415–435.

Russell CA, Fonville JM, Brown AE, Burke DF, Smith DL, James SL, Herfst S, Fouchier R, Smith DJ (2012) The potential for respiratory droplet-transmissible A/H5N1 influenza virus to evolve in a mammalian host. *Science* **336**: 1541–1547.

Schafer JR, Kawaoka Y, Bean WJ, Suss J, Senne D, Webster RG (1993) Origin of the pandemic 1957 H2 influenza A virus and the persistence of its possible progenitors in the avian reservoir. *Virology* **194**: 781–788.

Scholtissek C, Rohde W, VonHoyningen V, Rott R (1978) On the origin of the human influenza virus subtypes H2N2 and H3N2. *Virology* **87**: 13–20.

Sharp PM, Matassi G (1994) Codon usage and genome evolution. *Curr Opin Genet Dev* **4**: 851–860.

Shih AC, Hsiao TC, Ho MS, Li WH (2007) Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc Natl Acad Sci* **104**: 6283–6288.

Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Siu KM, Rambaut A (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic *Nature* **459**: 1122–1126; doi:10.1038/nature08182.

Suoeka N (1992) Directional mutational pressure, selective constraints, and genetic equlibria. *J Mol Evol* **34**: 95–114.

Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci* **101**: 11030–11035.

Than C, Nakhleh L (2009) Species tree inference by minimizing deep coalescences. *PLoS Comput Biol* **5**(9): e1000501. doi:10.1371/journal.pcbi.1000501.

Vijaykrishna D, Smith GJD, Pybus OG, Zhu H, Bhatt S, Poon LL, Peiris JM (2011) Long-term evolution and transmission dynamics of swine influenza A virus. *Nature* **473**: 519–523; doi:10.1038/nature10004.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2 — a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191; doi:10.1093/bioinformatics/btp033.

Whitehead TA, Fleishman SJ, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Baker D (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332**: 816–821.

Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, Baker D (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature Biotechnol* **30**: 543–548.

Wong HME, Smith DK, Rabadan R, Peiris M, Poon LL (2010) Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evolutionary Biology* **10**: 253; http://www.biomedcentral.com/1471-2148/10/253.

Zama M (1990) Codon usage and secondary structure of mRNA. *Nucl Acids Symp* **22**: 93–94.