

Novel algorithm for phylogenetic analysis of proteins: application to analysis of the evolution of H5N1 influenza viruses

Vladimir R. Perovic

Received: 27 October 2012 / Accepted: 11 June 2013 / Published online: 22 June 2013
© Springer Science+Business Media New York 2013

Abstract The highly pathogenic avian influenza virus (HPAIV) A subtype H5N1 is causing threat to human health over the years. Phylogenetic analysis is an important tool for analyzing the evolution of influenza. A novel phylogenetic algorithm based on a new protein distance measure derived from the informational spectrum method (ISM) has been presented. The new phylogenetic approach allows assessment of functional evolution of protein sequences. The new ISM-based phylogenetic approach has been found to overcome some drawbacks of other phylogenetic approaches, particularly concerning sensitivity to a single mutation, deletion and the position of the mutation. The ISM-based approach applied to hemagglutinin subunit 1 protein (HA1) of HPAIV A subtype H5N1 viruses in Egypt between 2006 and 2011, revealed clear clustering in two groups, with one growing group of H5N1 viruses after 2009 with increased number of human infections with H5N1. Four group-specific mutations are identified which are important for increased human tropism and the pandemic potential.

Keywords Protein sequence · Phylogenetic analysis · H5N1 influenza virus · Electron-ion interaction potential · Informational spectrum method

1 Introduction

The threat from pandemic influenza has been increasingly highlighted over the last few years as the H5N1 highly pathogenic avian influenza virus (HPAIV) has continued

Electronic supplementary material The online version of this article (doi:[10.1007/s10910-013-0212-1](https://doi.org/10.1007/s10910-013-0212-1)) contains supplementary material, which is available to authorized users.

V. R. Perovic (✉)
Institute of Nuclear Sciences Vinca, Center for Multidisciplinary Research,
University of Belgrade, P.O. Box 522, 11001 Belgrade, Serbia
e-mail: vladaper@vinca.rs

to spread and human infections have continued to occur. HPAIV, which has been transmitted from birds to humans on a limited basis, can be rapidly fatal in humans, with an overall case fatality rate of $>50\%$ [1]. If a human-to-human transmissible form emerges and spreads rapidly, it will pose a great threat to global public health, although the mortality rate may not be as high as that observed thus far. Among all human H5N1 cases reported worldwide between 2009 and 2011, 58.6% occurred in Egypt [2], providing ample opportunities to further adapt to humans in this country. While various factors suggest that this strain could be the cause of the next pandemic, it is unknown when it will occur, whether it will be caused by H5N1 or another new virus or how severe it will be. The development of new approaches, which will help to understand the significance of structural changes resulting from the alarming mutational propensity for human-to-human transmission of HPAIV, is of particular interest.

The phylogeny, or the evolutionary history of a particular species, is an important component when analyzing the evolution of influenza. Phylogenetic analysis can help determine past viruses and their patterns as well as determining a common ancestor of the virus. Past studies revealed that an avian virus spread to pigs and then to humans approximately 100 years ago [3]. This resulted in human lineages further evolving and becoming more prominent and stable. Phylogenetic approaches which are currently applied for analysis of the evolution of influenza viruses are based on the multiple sequence alignment (MSA). These approaches, although give some useful information about the evolution of influenza viruses, also have some serious drawbacks. The main weaknesses of the current phylogenetic algorithms for analysis of protein sequences are (i) insensitivity to position of the mutations, and (ii) failure to consider deletion within sequence. Recently reported results demonstrated that positions and the type of substituted amino acids are crucial for infection of humans by H5N1 virus. Of note also is that single deletion S129 in hemagglutinin from H5N1 virus significantly increases its human tropism [4,5].

In order to overcome these drawbacks of the phylogenetic analysis of influenza viruses, a novel algorithm is developed which is based on the informational spectrum (IS) representing information encoded in the protein primary structure. Here proposed phylogenetic algorithm is sensitive to the position and deletion and allows analysis of the functional evolution of proteins. Phylogenetic trees of HA from Egyptian H5N1 viruses, extensively evolving towards the human interacting profile, are generated based on informational spectrum method (ISM), conventional methods, and compared.

2 Materials and methods

2.1 Informational spectrum method

The basis of the ISM is the hypothesis that the protein-protein interaction includes two basic steps: (i) recognition and targeting between interacting proteins for long-range interactions at distances $>100\text{ \AA}$ and (ii) chemical binding for short range interactions at distances $<5\text{ \AA}$. The long-range properties of biological molecules are determined by the electron-ion interaction potential (EIIP). EIIP represents the main energy term

Table 1 The EIIP used to encode the amino acids

Amino acid	EIIP (Ry)
Leu	0.0000
Ile	0.0000
Asn	0.0036
Gly	0.0050
Glu	0.0057
Val	0.0058
Pro	0.0198
His	0.0242
Lys	0.0371
Ala	0.0373
Tyr	0.0516
Trp	0.0548
Gln	0.0761
Met	0.0823
Ser	0.0829
Cys	0.0829
Thr	0.0941
Phe	0.0946
Arg	0.0959
Asp	0.1263

of valence electrons [6], and for organic molecules it can be calculated using the following equation derived from the “general model pseudopotential” [7,8]:

$$W = 0.25 \frac{Z^* \sin(1.04\pi Z^*)}{2\pi} \quad (1)$$

where Z^* is the average quasivalence number (AQVN) defined by

$$Z^* = \frac{1}{N} \sum_{i=1}^m n_i Z_i \quad (2)$$

where m is the number of atomic components in the molecule, Z_i is the valence number of the i th atomic component, n_i is the number of atoms of the i th component and N is the total number of atoms. The EIIP values calculated according to Eqs. (1) and (2) are expressed in Rydbergs (Rys) units.

The ISM starts by assigning the EIIP value (Table 1) to each amino acid of the protein sequence. This new numerical sequence, corresponding to the protein sequence, is then subjected to a discrete Fourier transformation which is defined as follows:

$$X(n) = \sum_{m=1}^N x(m)e^{-i2\pi n(m-1)/N}, \quad n = 1, 2, \dots, N/2 \quad (3)$$

where $x(m)$ is the m th member of a given numerical series, $X(n)$ is the n th coefficient discrete Fourier transformation and N is the total number of points in this series. These coefficients describe the amplitude, phase and frequency of sinusoids which constitute the original signal. The complete information about the original sequence is contained in both amplitude and phase spectra. However, in the case of protein analysis, relevant information is presented in an energy density spectrum [9], which is defined as follows:

$$S(n) = X(n)X^*(n) = |X(n)|^2, \quad n = 1, 2, \dots, N/2 \quad (4)$$

In this way, sequences are analyzed as discrete signals. It is assumed that their points are equidistant with the distance $d = 1$. The maximal frequency in a spectrum defined as above is $F = 1/2d = 0.5$. The frequency range is independent of the total number of points in the sequence. The total number of points in a sequence influences only the resolution of the spectrum. The resolution of the sequence of length N is $f = 1/N$. The n th point in the spectrum corresponds to a frequency $f(n) = n f = n/N$. Thus, the initial information defined by the sequence of amino acids can now be presented in the form of the IS, representing series of frequencies and their amplitudes.

2.2 Phylogenetic analysis

Phylogeny defines the evolutionary relationship within a family of closely related sequences. Methods for constructing and analyzing phylogenetic trees can be classified into distance-based and character-based methods [10, 11]. The distance-based methods first calculate distances between each pair of sequences according to various measures and evolutionary models [12–15]. In the second step, the distance matrix is transformed into a tree using some of the clustering algorithms e.g. unweighted pair group method with arithmetic mean (UPGMA) [16, 17], the neighbor-joining (NJ) [18] and Fitch–Margoliash method [19]. The character-based methods find the best tree evaluating the fitness among different topologies, where the calculations are carried out on each of the individual residues of sequences. Based on the optimality criterion character-based methods can be classified into (i) maximum parsimony methods (MPs) [20–22], with the strategy to minimize the number of evolutionary changes e.g. substitutions, (ii) methods based on a probabilistic approach to phylogeny which includes maximum likelihood methods (MLs) [23] and Bayesian inference [24] based on Monte–Carlo sampling [25]. Depending on the level of sequence similarity, maximum parsimony methods are often used for quite similar sequences, distance-based methods for sequences that still share recognizable sequence similarity, and probabilistic methods for group of sequences with lower level of similarity [26].

Distance-based methods and their implementations (Fitch, Kitsch, Neighbor tools of the Phylip package [27], ClustalW and X [28], MEGA [29]) and MP methods (Protpars [27], PAUP [30]) are relatively fast and can handle a large group of sequences,

whereas probabilistic techniques (ProtML [31], Tree-PUZZLE [32], PAML [33], MrBayes [34], BEAST [35], PHYML [36]) are more accurate but with large computational cost [37–39]. Recently novel methods are developed based on discrete mathematical approach e.g. Boolean analysis or BOOL-AN [40].

All of these approaches require MSA and most of them assume evolutionary models and empirical substitution matrices. The quality of MSA implies the quality of inferred phylogenetic tree [41]. The problems with MSA are time complexity, limited number of sequences, choosing the parameters for MSA programs [39,42], controversial evolutionary models, ambiguity of the alignment cost criteria. Different tools produce different alignments [43]. In the popular progressive MSA programs (ClustalW [44], T-Coffee [45], MAFFT [46]) the reliance of MSA is connected to the alignment of the first two most closely related sequences [44]. To overcome those problems during the last few years several alignment-free methods for phylogenetic analysis of protein sequences have been developed, based on different sequence distances e.g. Bhattacharyya distance [47], Lempel-Ziv complexity [48,49], using feature vectors [50], relative complexity measure [42,51], and the method for highly divergent sequences PHYRN [39].

2.3 A new protein distance measure based on ISM

The important flaw of the MSA-based phylogenetic analyses is that sequence similarity does not automatically imply similarity in biological functions. For example, two protein sequences that differ by a single mutation that is lethal for the biological function will be phylogenetically close, whereas two proteins that differ in several mutations that do not affect biological functions will be phylogenetically separated. To overcome this drawback and to improve functional sequence analysis, a new MSA-independent protein distance measure, based on the ISM is proposed.

Let X and Y be two sequences, let $S_x = \{S_x(n)\}$ and $S_y = \{S_y(n)\}$, $n = 1, 2, \dots, N/2$, be their corresponding energy density spectra, where N is length of the longest sequence. Then the distance between X and Y is defined as:

$$d(X, Y) = \frac{1}{N} \sum_{n=1}^{N/2} |S_x(n) - S_y(n)| \quad (5)$$

2.4 Properties of the new distance

- (i) Distance d defined in (5) is a Minkowski L_1 distance (multiplied by a factor $1/N$) on vector space $\mathbf{R}^{N/2}$ of informational spectra. As a result of the Minkowski inequality [52] it is a valid metric measure:

It satisfies:

- 1) $d(x, y) \geq 0$, non-negativity
- 2) $d(x, y) = 0 \Leftrightarrow x = y$, identity of indiscernibles
- 3) $d(x, y) = d(y, x)$, symmetry
- 4) $d(x, z) \leq d(x, y) + d(y, z)$, triangle inequality

- (ii) ISM phylogenetic approach is not based on MSAs and does not use any of the substitution models.
- (iii) The ISM distance d is sensitive to the position of mutation and the type of the substituted residue, in opposition to standard approaches where all sites are treated equally, like in Dayhoff [53] and Jones–Taylor–Thornton (JTT) [54] substitution models where protein distance is sensitive only to the type of mutation. The same amino acid substitution in different sites implies changing of the EIIP values in the same different positions of the sequence signal vector $x(m)$, which then implies different changing in the whole informational spectra $S(n)$ as defined in (3) and (4) of mutated sequence. Therefore distances d defined in (5) between non-mutated and mutated sequence directly depends of the position of single mutation. In Fig. 1 and Table 2 are given the average and standard variation of the ISM distances and JTT distances between non-mutated and mutated sequences as a function of substitution over all of 110 amino acid positions in Homo sapiens insulin sequence with each of I, A and D amino acids, representing the minimum, medium and maximum EIIP values.
- Because of the last two properties (ii) and (iii) no conventional validation approaches (e.g. bootstrap or likelihood) can be applied in ISM-based phylogenetic approach.
- (iv) The sensitivity to a single mutation and position of the mutation. The set of 110 mutated sequences was obtained by introducing single mutation with amino acid I in each one of the 110 amino acid positions in Homo sapiens insulin sequence. With the same procedure two more sets were generated with mutation with amino acids A and D. Phylogenetic trees with standard and ISM-based algorithm were generated for those sets and compared. The trees generated by ISM-based phylogenetic approach (Fig. 2c, f, i) show more diversity and sensitivity to single mutation position, compared to the trees obtained by standard phylogenetic approach (Fig. 2a, b, d, e, g, h).
- (v) Calculation of EIIP for organic molecules, including amino acids, is based only on the bruto chemical formula. It means that EIIP values for the two isomers of some organic molecule are the same. Therefore Leucine and Isoleucine have the same value of the EIIP which are equal to 0 Ry. According to the ISM concept, replacement of residues in some protein with amino acids with the same EIIP values (in the particular case Leu \leftrightarrow Ile) will not affect its IS characteristic representing the long-range properties of the protein. In other words, such mutations do not change the position of the protein in the ISM-based phylogenetic tree.

2.5 Algorithm

The ISM-based phylogenetic tree was generated using the following algorithm:

1. For each sequence calculate its spectrum:
 - 1.1 Convert amino acid sequence into signal with EIIP values.
 - 1.2 Decrease signal to zero mean.

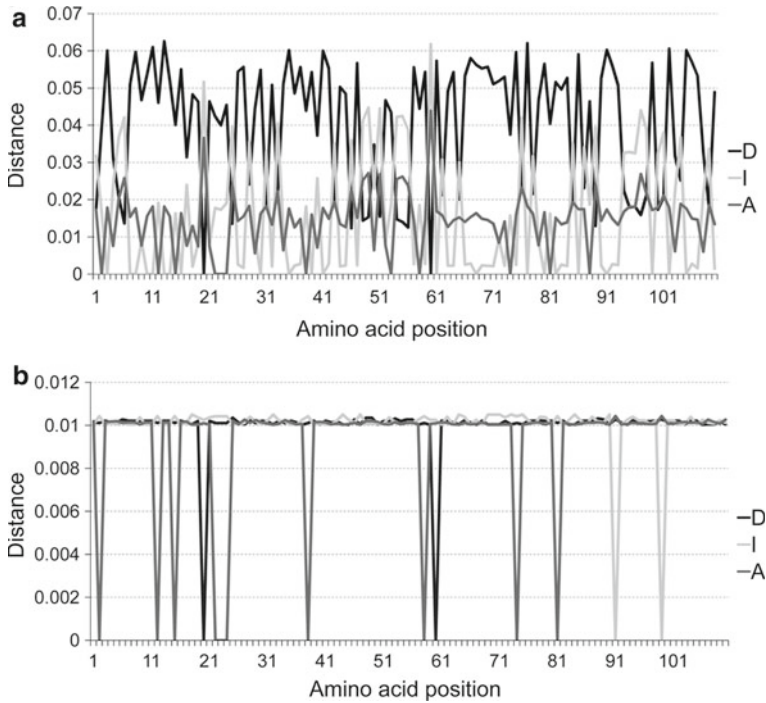


Fig. 1 The graphic presentation of variation of the **a** ISM-based distance and **b** JTT distance between protein sequence of human insulin and mutated sequences as a function of single mutation over all 110 amino acid positions of human insulin (GenBank: AAA59172.1) with I, A and D

Table 2 Average values and variations of the ISM and JTT distances between insulin Homo sapiens and mutated sequences corresponding to single mutations over all 110 amino acid positions with I, A and D

Distance	Amino acid	Average	Standard deviation	Coefficient of variation
ISM	I	0.015968987	0.016462359	1.030895669
	A	0.014680109	0.007753097	0.528136207
	D	0.040545540	0.017047781	0.420460072
JTT	I	0.010035816	0.001381754	0.137682230
	A	0.009192425	0.002920772	0.317736800
	D	0.009981375	0.001367115	0.136966642

1.3 Zero-padding to length of the longest signal, to set the same resolution to all spectra.

1.4 Apply Fast Fourier Transformation to signal to generate energy density spectrum.

2. Calculate the distance matrix with the distance measure defined in (5).

3. Construct the tree using the UPGMA method.

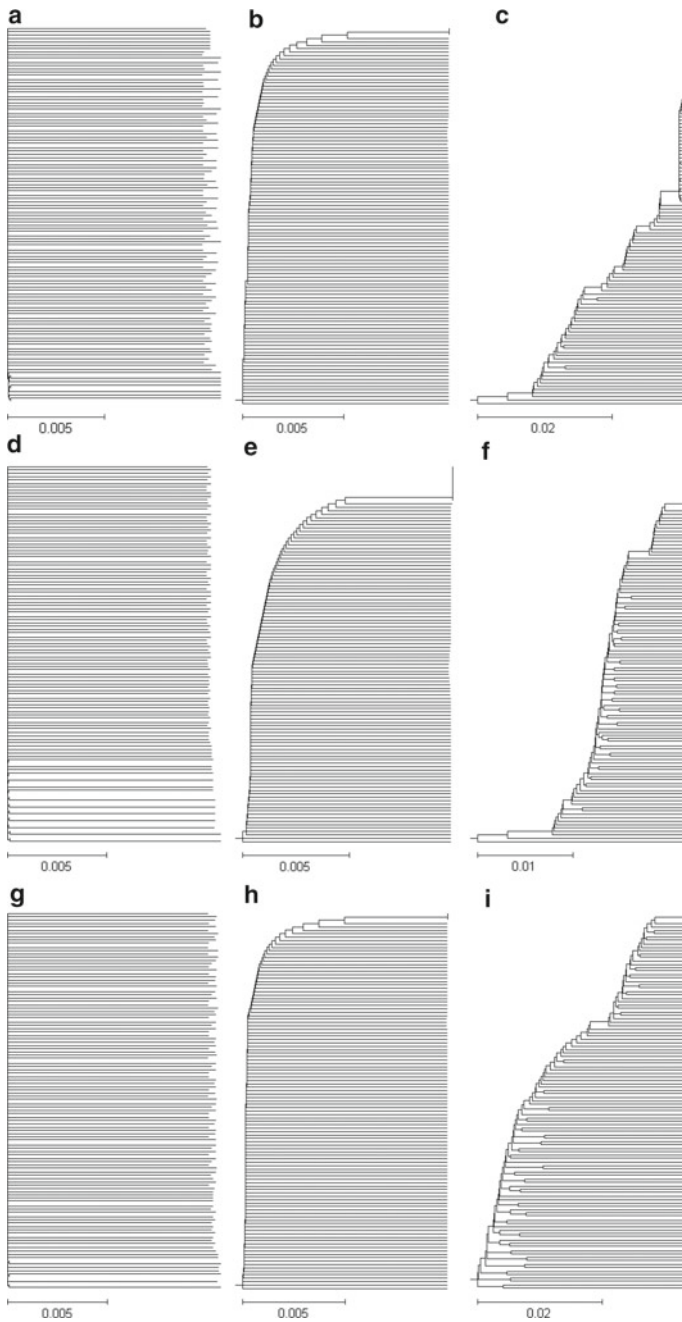


Fig. 2 Phylogenetic tree for each set of 110 sequences obtained by the introduction of the each selected amino acids (I, A, D) in 110 amino acid positions of human insulin protein. **a, d, g** phylogenetic trees constructed using the MSA-based NJ method, **b, e, h** MSA-based UPGMA method, **c, f, i** ISM-based algorithm

2.6 Algorithm complexity

The total time complexity for the algorithm of generating ISM-based tree is $O(NL(N + \log L))$, where N is the number of sequences and L is the length of the longest sequence. For the first step of the algorithm the time complexity is $O(NL \log L)$ according to the time complexity of Fast Fourier Transformation $O(L \log L)$ [55]. For the calculation of distance matrix in the second step the time complexity is $O(N^2L)$ and for the third step of the UPGMA clustering the complexity is $O(N^2)$ [56].

2.7 Computing time

Performance test of ISM-based phylogenetic method was carried out on simulated data sets. The computing times and used memory by ISTREE software depending on number and length of sequences are given in Table 3. The data sets of protein sequences were randomly generated. The number of sequences for generated data sets was in the range from 100 to 4,000, and the lengths of sequences were between 100 and 10,000. The computing time was measured on a PC Pentium[®] Dual-Core CPU E5200 2.50 GHz 3 GB RAM, running on Windows XP system. For the group of 4,000 sequences with the length of 10,000 residues the total computing time is 34 min 17 s. Detailed computing times for every phase of the method are given in Table 3 in format SP+DM+CL, where SP is a measured time for the first phase of calculating informational spectra, DM for the second phase of generating distance matrix and CL for the third phase of clustering.

Table 3 Computing times and used memory by ISMTREE software for randomly generated protein sequences. Computer: Pentium[®] Dual-Core CPU E5200 @ 2.50 GHz, 3 GB of RAM. System: Windows XP. Detail times for every phase are given in format SP+DM+CL, where SP is measured time for the first phase of calculating informational spectra, DM for the second phase of generating distance matrix and CL for the third phase of clustering

Number of sequences	Length of sequences	Total time	SP+DM+CL (s)	Used memory (MB)
100	100	0.11 s	0.001+0.094+0.015	1
	1,000	0.22 s	0.047+0.158+0.015	2
	10,000	2.844 s	1.610+1.219+0.015	9
1,000	100	11.06 s	0.06+9+2	18
	1,000	17.5 s	0.5+15+2	20
	10,000	2 min 16 s	16+118+2	74
2,000	100	50.12 s	0.12+35+15	35
	1,000	1 min 15 s	1.1+59+15	74
	10,000	9 min	27+498+15	193
4,000	100	4 min 22 s	0.26+141+121	254
	1,000	5 min 58 s	2.17+235+121	269
	10,000	34 min 17 s	53+1,880+121	499

2.8 Software

Conventional distance-based trees were inferred using MEGA5 [57] software package, and for the ML trees the PHYML [36] tool was applied. For the MSA calculation of sequences the MUSCLE algorithm [58] of MEGA5 software was used.

Web service *ISTREE* was developed for generating ISM-based trees. *ISTREE* is freely available on the web page <http://istree.bioprotection.org>.

2.9 Sequences

For comparison and testing the phylogenetic analyses, all completed and non-predicted sequences of glucocorticoid receptor, insulin, leptin hormone and lipoprotein lipase enzyme were downloaded from NCBI database [59]. For phylogenetic analysis of H5N1 HA1, all of the 526 published sequences of H5N1 influenza A viruses isolated in Egypt between 2006 and 2011 were downloaded from NCBI and GISAID databases [60] and submitted to analysis.

3 Results

3.1 Testing sets

For testing and comparing standard and ISM phylogenetic approaches, for each set of sequences of glucocorticoid receptor, insulin protein, leptin hormone protein and lipoprotein lipase enzyme, phylogenetic trees were generated using (i) ML method with NNIs tree topology search and BioNJ method for initial tree search, (ii) UPGMA method with Poisson correction model, where all ambiguous positions were removed for each sequence pair, and (iii) ISM approach with distance measure defined in (5). Comparison of the ISM-based and the standard phylogenetic methods shows similar biological classification of taxonomic classes and orders of species, but also reveals some differences (Fig. 3, Online Resource 1). These differences are not significant and concern the shift inside of biologically related branches e.g. only in the ISM based trees for glucocorticoid receptor and leptin protein all Rodentia species are grouped together (Fig. 3c, i); in the ISM and UPGMA trees of LPL Guinea pig are diverged from Rodentia (Fig. 3k, l); in ML tree Rabbit is misplaced in Primates branch (Fig. 3i) etc. All of the trees in Fig. 3 follow the standard classification to Eutheria (Placental Mammals), Marsupialia, Aves (Birds), Salientia and Teleostei (Fish).

Algorithms for standard molecular phylogenetic analysis, which use the information obtained by a structural analysis of defined set of genes in different species, are principally different from the ISM-based phylogenetic algorithm, analyzing the functional evolution of a single gene across multiple species. Comparison of the trees from Fig. 3 with the commonly accepted molecular phylogenetic trees of Placental Mammals [61–63] reveals similar clustering of Placental Mammals into Afrotheria, Eutheria, Laurasiatheria, with some differences: (i) For the glucocorticoid receptor (Fig. 3a–c) *Loxodonta Africana* (Afrotheria) is separated from

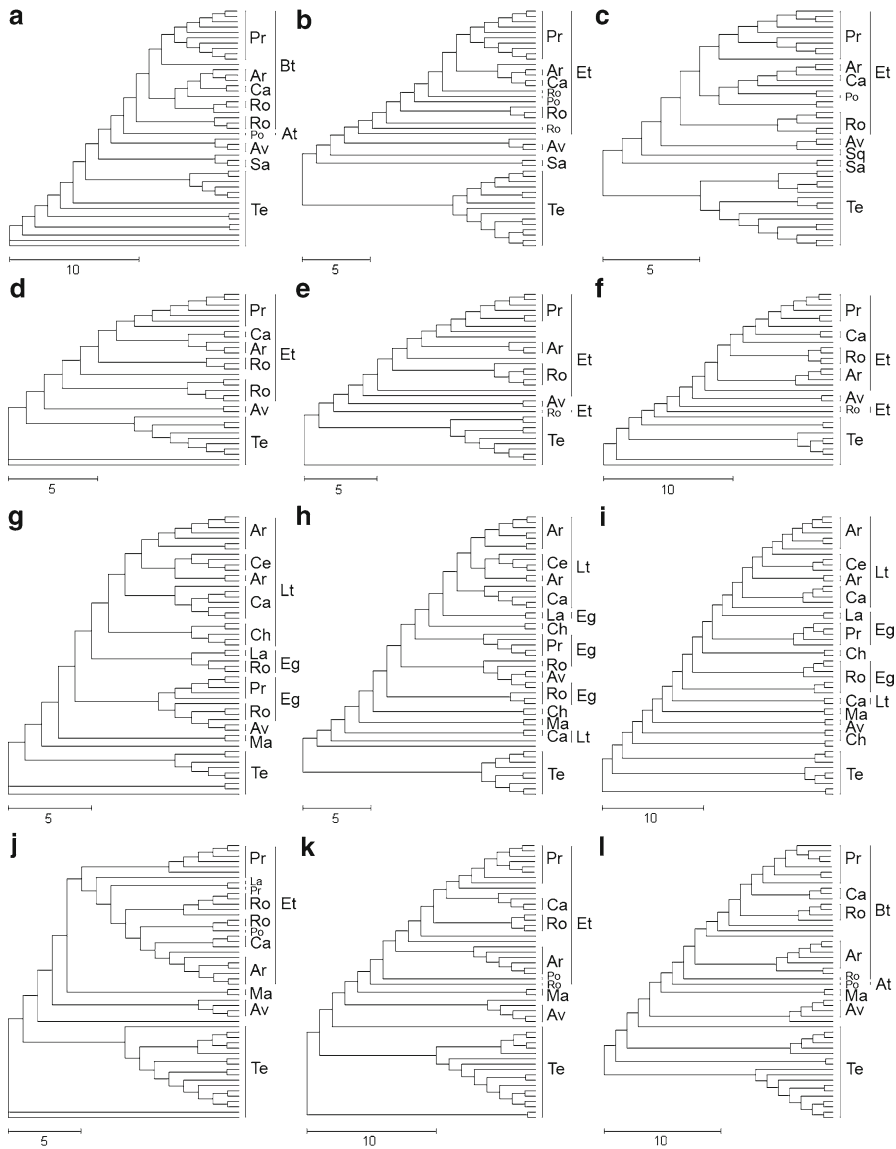


Fig. 3 Comparison of the standard and ISM-based phylogenetic analyses; Phylogenetic tree constructed using **a, d, g, j** standard ML method; **b, e, h, k** standard UPGMA method; **c, f, i, l** ISM-based method for **a, b, c** glucocorticoid receptor; **d, e, f** insulin protein; **g, h, i** leptin hormone; **j, k, l** lipoprotein lipase enzyme. Abbreviations for taxonomic classes and orders of species: *Ar* Artiodactyla, *Av* Aves (Birds), *Ca* Carnivora, *Ce* Cetacea, *Ch* Chiroptera, *La* Lagomorpha, *Ma* Marsupialia, *Pr* Primates, *Po* Proboscidea, *Ro* Rodentia, *Sa* Saliencia, *Sc* Scandentia, *Sq* Squamata, *Te* Teleostei (Fish). Abbreviations for clades: *At* Afrotheria, *Bt* Boreoeutheria, *Eg* Euarchontoglires, *Et* Eutheria, *Lt* Laurasiatheria

Boreoeutheria cluster only in the ML tree (Fig. 3a). (ii) In the case of insulin protein (Fig. 3d–f), all Eutheria species are clustered together in the ML tree, while in the UPGMA and ISM trees a few Rodentia (Degu and Naked mole rat) are

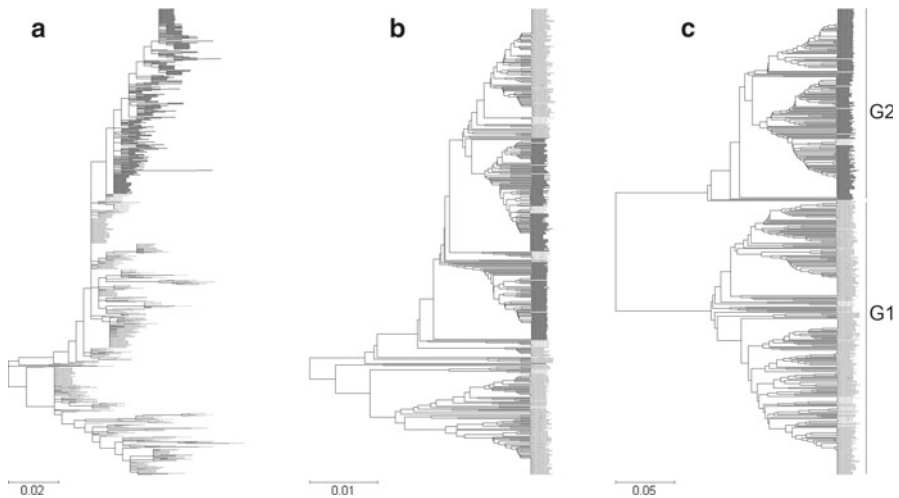


Fig. 4 Phylogenetic analysis of H5N1 viruses isolated in Egypt between 2006 and 2011; Phylogenetic trees are constructed using: **a** the ML method; **b** MSA-based UPGMA method; **c** ISM-based method. HA1 Sequences with G1 specific residues D43, S120, (S,L)129, I151 are colored in light grey, and sequences with G2 specific residues N43, (D,N)120, 129del, T151 in dark grey (Table 5)

separated from Eutheria branch. (iii) The trees of leptin protein (Fig. 3g–i) show the most similarities to the standard molecular phylogenetic tree, but a few Chiropteras in the ISM and UPGMA trees are divided from Eutheria cluster (Fig. 3h, i), and in UPGMA tree (Fig. 3h) Aves are misclustered in Eutheria lineage. (iv) In the LPL trees (Fig. 3j–l) Afrotheria is separated from Boreoeutheria only in the ISM tree (Fig. 3l).

3.2 Analysis of Egyptian H5N1 HA1

For the set of all published H5N1 HA1 sequences from Egypt between 2006 and 2011, phylogenetic trees were generated using the standard approaches and the ISM-based algorithm as described in Sect. 2 (Fig. 4, Online Resource 2). The ISM-based phylogenetic tree shows clear clustering in two distinct groups G1 and G2. Some similar groupings also appear in standard trees but are not clearly separated as clustered in ISM-based tree (Fig. 4).

Detailed ISM-based phylogenetic analysis of Egyptian H5N1 viruses for each year between 2006 and 2011 and the three year periods 2006–2008 and 2009–2011 (Fig. 5), shows that number of sequences in G2 (Table 4; Fig. 6) is in steady increase from 0% in 2006 (Fig. 6a) to 95.95% in 2011 (Fig. 6g). Moreover, the analysis reveals sudden increase in number of viruses in G2 after 2008, from 6.06% in 2008 to 54.55% in 2009, and from 6.25% in period 2006–2008 to 41.25% in period 2009–2011 (Fig. 6d, h), which is in correlation to the sudden increase in the number of human cases of avian influenza H5N1 after 2008, confirmed by World Health Organization (WHO) [64].

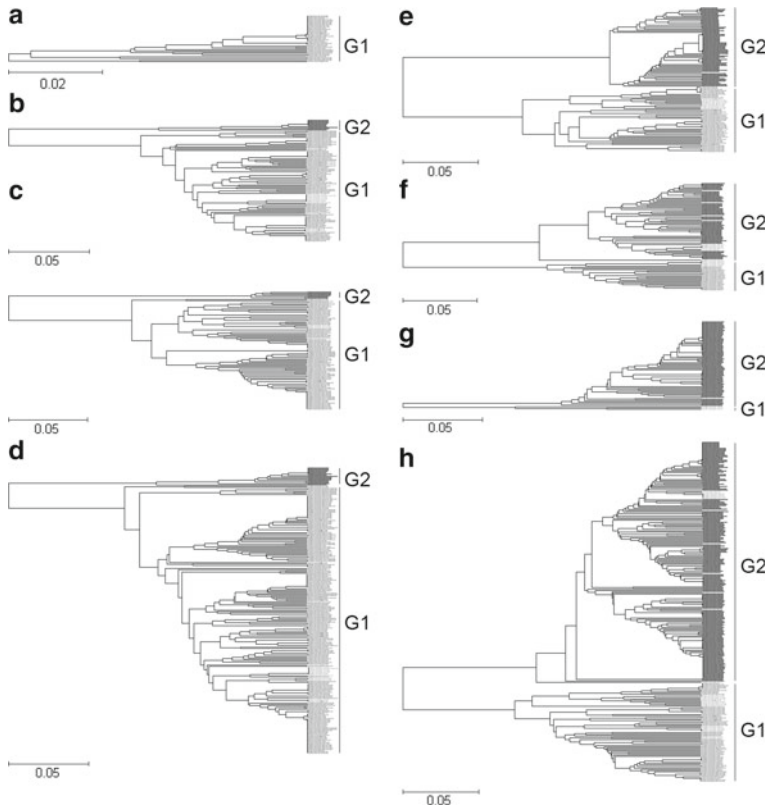


Fig. 5 Detailed ISM-based phylogenetic analysis of Egyptian H5N1 viruses for each year between 2006 and 2011; ISM-based tree for: **a** 2006; **b** 2007; **c** 2008; **d** period 2006–2008; **e** 2009; **f** 2010; **g** 2011; **h** period 2009–2011

Table 4 Distribution of Egyptian H5N1 viruses between G1 and G2 groups in ISM-based trees constructed for every year between 2006 and 2008, for 3 year periods 2006–2008, 2009–2011 and the whole period 2006–2011

Year	G1	G2
2006	39 (100 %)	0 (0 %)
2007	93 (91.18 %)	9 (8.82 %)
2008	93 (93.94 %)	6 (6.06 %)
2006–2008	225 (93.75 %)	15 (6.25 %)
2009	55 (45.45 %)	66 (54.55 %)
2010	25 (27.78 %)	65 (72.22 %)
2011	3 (4.05 %)	71 (95.95 %)
2009–2011	83 (29.12 %)	202 (70.88 %)
2006–2011	309 (58.75 %)	217 (41.25 %)

Homology analysis of differences between G1 and G2 revealed four group-specific amino acid positions (Table 5): D43, S120, (S, L) 129, I151 in G1 and N43, (D, N) 120, 129del, T151 in G2.

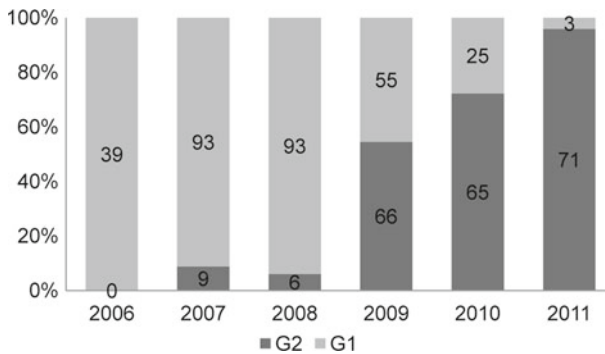


Fig. 6 Distribution of Egyptian H5N1 viruses between G1 and G2 groups in ISM-based trees constructed for every year between 2006 and 2008 (Table 4)

Table 5 Residues with their percentage of occurrence, which are specific for H5N1 HA1 in groups G1 and G2

	Group G1	Group G2
Group-specific residues	D43 (99%)	N43 (98%)
	S120 (94%)	(D,N)120 (94%)
	(S,L)129 (98%)	129del (99%)
	I151 (92%)	T151 (99%)

For comparing the sensitivity of ISM-based and standard phylogenetic approaches to mutations and deletions, all of the 309 sequences from G1 are selected as a test set. Half of them (every second (s)) are mutated by introducing characteristic mutations of G2 (43N, 120D, 129del, 151T). In opposition to the standard phylogenetic approach, the ISM-based phylogenetic tree reveals clear clustering and separation of mutated and non-mutated sequences (Fig. 7).

4 Discussion

Modulation of the biological properties of proteins by mutations depends on their positions and the type of the substituted amino acids. The main weaknesses of the phylogenetic approaches which are based on the MSA are their insensitivity to these important changes in the primary structure of proteins. In order to test sensitivity of here proposed ISM-based phylogenetic algorithm to position and the type of mutations the primary structure of human insulin was scanned by replacing all residues with amino acids I, A and D, representing the minimum, medium and maximum EIIP values of the EIIP range (Table 1). In Fig. 1 are given the distance values corresponding to mutations in each position in the primary structure of human insulin. As can be seen, the presented distance values strongly depend on the position of mutation and the type of the substituted residue.

For further comparison of the sensitivity of the sequence similarity-based and the ISM-based phylogenetic algorithms to the mutations, three sets of proteins were used. Each set contained 110 proteins and was obtained by substitution

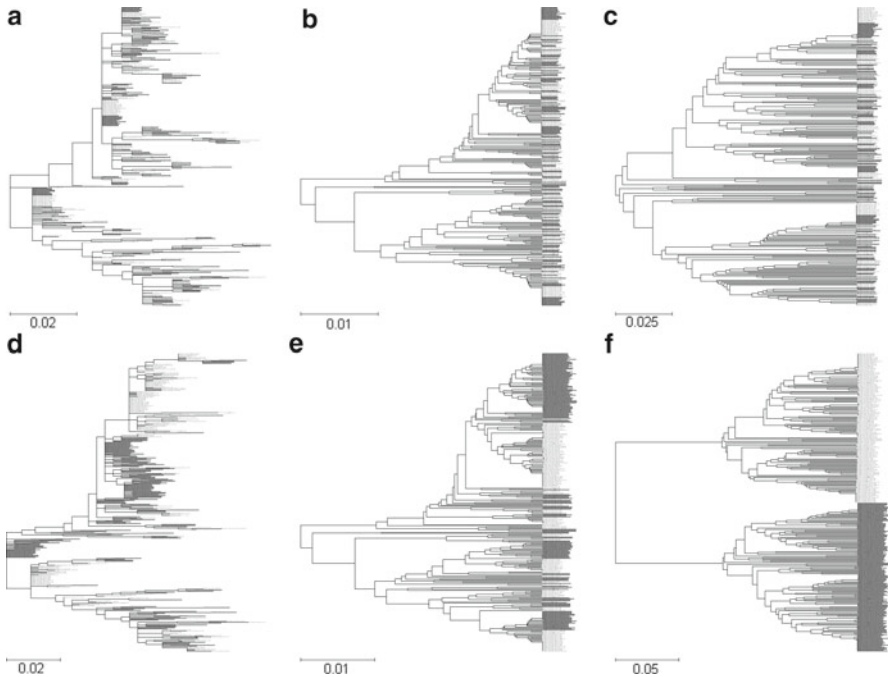


Fig. 7 Comparison of sensitivity of standard MSA-based and ISM-based phylogenetic algorithms for the detection of mutations which are important for human tropism of Egyptian H5N1 viruses; **a–c** phylogenetic trees constructed for all 309 unmutated G1 sequences; **d–f** trees for all G1 sequences where each second sequence was mutated with D43N, S120D, S129del and I151T. Trees are constructed using: **a, d** standard ML method; **b, e** standard UPGMA method; **c, f** ISM-based algorithm. HA1 sequences selected for mutations are colored in *dark grey*

residues in all 110 positions of the primary structure of human insulin by I, A or D. Phylogenetic trees constructed using the MSA-based NJ and UPGMA methods, and ISM-based method are presented in Fig. 2. According to the presented results, sensitivity to the position and the type of mutation is remarkable higher in the ISM-based method than in the MSA-based NJ and the MSA-based UPGMA methods.

As a further proof-of-concept of the here presented novel phylogenetic algorithm, the phylogenetic trees of glucocorticoid receptor, insulin, leptin and lipoprotein lipase from different species, constructed by the ISM-based and the MSA-based methods were compared. Results presented in Fig. 3 show that both methods presented the evolution of these four proteins in a similar way.

Currently, highly pathogenic avian influenza (HPAI) virus H5N1, causing high mortality in humans (mortality rate >60%), represents one of most serious threats for human population. Although H5N1 viruses still lack the ability to transmit efficiently among humans, this obstacle may be overcome by reassortment with cocirculating human H1N1 or H3N2 influenza viruses. On the other hand, the pandemic (H1N1) 2009 influenza virus that has spread globally in humans after transmission from the zoonotic reservoir, or the novel H3N2v influenza virus which start to spread in US, could potentially acquire the high lethality of the H5N1 HPAI virus for humans via

reassortment. The HPAI virus H5N1 emerged in Egypt in 2006 and was declared endemic in 2008. Among all human H5N1 cases reported worldwide between 2009 and 2011, 58.6 % occurred in this country alone. Since Egypt became the new epicenter of human H5N1 infections, the increasing pandemic potential of H5N1 is a major concern.

In Fig. 4 are presented the phylogenetic trees of 526 hemagglutinin subunit 1 (A1) sequences from H5N1 viruses isolated in Egypt between 2006 and 2011, constructed by the ISM-base method and the MSA-based ML and UPGMA methods. As can be seen, in the tree constructed by ISM-based method, these HA1 sequences form two clearly divided subgroups denoted by G1 and G2. This clustering is also visible, although less sharp, in other two presented trees constructed by standard ML and UPGMA methods. Sequence homology analysis revealed that HA1 89 % of G1 viruses had the amino acids D43, S120, (S,L)129 and I151, whereas 94 % of G2 viruses had N43, (D,N)120, 129 Δ and T151 (amino acid positions according to H5 numbering). Both in poultry and human viruses the four G2-typical amino acids steadily increased with time and were mostly acquired after 2008. Of note is that the number of human infections by H5N1 viruses in Egypt significantly increased from 2009, suggesting that these four mutations were important for increased human tropism and the pandemic potential of these viruses.

Finally, the mutations N43, (D, N) 120, 129 Δ and T151, which characterize G2 viruses, were introduced in the half of 309 HA1 from G1 viruses, and the trees by the ISM-based method and standard ML and UPGMA methods were constructed for these mutated and unmutated sequences (Fig. 7). In contrast to the trees obtained by MSA-based methods (Fig. 7d, e), the mutated HA1 sequences in the tree constructed by ISM-based method form well defined cluster (Fig. 7f). This result represents an additional proof of the advantage of the ISM-based method over the MSA-based methods in terms of sensitivity to mutations which are essential for biological role of proteins.

In conclusion, here presented novel ISM-based phylogenetic method, in contrast to other conventional methods, is sensitive to positions and type of mutations and allows assessment of functional evolution of protein sequences.

Acknowledgments This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant No. 173001).

References

1. T.T. Wang, M.K. Parides, P. Palese, *Science* **335**(6075), 1463 (2012)
2. Influenza at the Human-Animal Interface (HAI). (World Health Organization) http://www.who.int/influenza/human_animal_interface/en/. Accessed June 2012.
3. C. Scholtissek, *Virus Genes*. **11**(2–3), 209–15 (1995)
4. Y. Watanabe, M.S. Ibrahim, H.F. Ellakany, N. Kawashita, R. Mizuike, H. Hiramatsu, N. Sriwilaijaroen, T. Takagi, Y. Suzuki, K. Ikuta, *PLoS Pathog.* **7**(5), e1002068 (2011)
5. V. Veljkovic, N. Veljkovic, C.P. Muller, S. Müller, S. Glisic, V. Perovic, H. Köhler, *BMC Struct. Biol.* **9**, 21 (2009)
6. V. Veljkovic, *A Theoretical Approach to Preselection of Cancerogens and Chemical Carcinogenesis* (Gordon & Breach, New York, 1980)
7. V. Veljkovic, I. Slavic, *Phys. Rev. Lett.* **29**, 105–107 (1972)

8. V. Veljkovic, Phys. Lett. A **45**, 41–42 (1973)
9. V. Veljkovic, I. Cosic, B. Dimitrijevic, D. Lalovic, IEEE Trans. Biomed. Eng. **32**, 337–341 (1985)
10. D. Swofford, G. Olsen, P. Waddell, D.M. Hillis, in *Molecular Systematics*, 2nd edition, ed. by D.M. Hillis, C. Moritz, B.K. Mable. (Sinauer, Sunderland, Massachusetts, 1996), pp. 407–511.
11. M.A. Hershkovitz, D.D. Leippe, in *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, vol. 39, ed. by A.D. Baxevanis, B.F.F. Ouellette (Wiley, Hoboken, NJ, USA, 2006). doi:10.1002/9780470110607.ch9.
12. T.H. Jukes, C.R. Cantor, *Mammalian Protein Metabolism* (Academic Press, New York, 1969), pp. 21–132
13. M. Kimura, J. Mol. Evol. **16**, 111–120 (1980)
14. H. Kishino, M. Hasegawa, J. Mol. Evol. **29**, 170–179 (1989)
15. J.A. Lake, Proc. Natl Acad. Sci. USA **91**, 1455–1459 (1994)
16. R. Sokal, C. Michener, Univ. Kansas Sci. Bull. **38**, 1409–1438 (1958)
17. P.H. Sneath, R.R. Sokal, *Numerical Taxonomy* (W. H. Freeman, San Francisco, 1973)
18. N. Saitou, M. Nei, Mol. Biol. Evol. **4**, 406–425 (1987)
19. W.M. Fitch, E. Margoliash, Science **155**, 279–284 (1967)
20. W.M. Fitch, Syst. Zool. **35**, 406–416 (1971)
21. D. Sankoff, R.J. Cedergren, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (Addison-Wesley, London, 1983)
22. J. Camin, R. Sokal, Evolution **19**, 311–326 (1965)
23. J. Felsenstein, Syst. Biol. **22**(3), 240–249 (1973)
24. B. Larget, D.L. Simon, Mol. Biol. Evol. **16**(6), 750 (1999)
25. W.K. Hastings, Biometrika **57**(1), 97–109 (1970)
26. W.M. David, *Bioinformatics: Sequence and Genome Analysis* (CSHL Press, New York, 2001)
27. J. Felsenstein, Cladistics **5**, 164–166 (1989)
28. J.D. Thompson, T. Gibson, D.G. Higgins, Curr. Protoc. Bioinf. 2–3 (2002)
29. S. Kumar, M. Nei, J. Dudley, K. Tamura, Briefings Bioinf. **9**(4), 299–306 (2008)
30. D.L. Swofford, B. Documentation, *Phylogenetic Analysis Using Parsimony* (Illinois Natural History Survey, Champaign, 1989)
31. J. Adachi, M. Hasegawa, in *MOLPHY Version 2.3: Programs for Molecular Phylogenetics Based on Maximum Likelihood*, No. 28. (Tokyo, Japan: Institute of Statistical Mathematics, 1996).
32. H.A. Schmidt, K. Strimmer, M. Vingron, A. von Haeseler, Bioinformatics **18**(3), 502–504 (2002)
33. Z. Yang, Mol. Biol. Evol. **24**(8), 1586–1591 (2007)
34. F. Ronquist, J.P. Huelsenbeck, Bioinformatics **19**(12), 1572–1574 (2003)
35. A.J. Drummond, A. Rambaut, BMC Evol. Biol. **7**(1), 214 (2007)
36. S. Guindon, O. Gascuel, Syst. Biol. **52**(5), 696–704 (2003)
37. T.L. Williams, B.M. Moret, in *Bioinformatics and Bioengineering, 2003. Proceedings.* (Third IEEE Symposium on, 2003), pp. 79–86.
38. S. Roch, Science **327**(5971), 1376–1379 (2010)
39. G. Bhardwaj, K.D. Ko, Y. Hong, Z. Zhang, N.L. Ho, S.V. Chintapalli, L.A. Kline, M. Gotlin, D.N. Hartranft, M.E. Patterson, F. Dave, E.J. Smith, E.C. Holmes, R.L. Patterson, D.B. van Rossum, PLoS One **7**(4), e34261 (2012)
40. É. Jakó, E. Ari, P. Ittész, A. Horváth, J. Podani, Mol. Phylogenet. Evol. **52**(3), 887 (2009)
41. L. Brocchieri, Theor. Popul. Biol. **59**(1), 27–40 (2001)
42. A. Albayrak, H.H. Otu, U.O. Sezerman, BMC Bioinf. **11**(1), 428 (2010)
43. X. Chen, M. Tompa, Nat. Biotechnol. **28**(6), 567–572 (2010)
44. J.D. Thompson, D.G. Higgins, T.J. Gibson, Nucl. Acids Res. **22**(22), 4673–4680 (1994)
45. P. Di Tommaso, S. Moretti, I. Xenarios, M. Orobitz, A. Montanyola, J.M. Chang, J.F. Taly, C. Notredame, Nucl. Acids Res. **39**(2), W13–W17 (2011)
46. K. Katoh, K.I. Kuma, H. Toh, T. Miyata, Nucl. Acids Res. **33**(2), 511–518 (2005)
47. S. Zhang, T. Wang, in *Proceedings of the 10th WSEAS International Conference on Applied Computer Science.* (WSEAS, 2010), pp. 67–71.
48. S. Zhang, T. Wang, MATCH Commun. Math. Comput. Chem. **63**(3), 701–716 (2010)
49. H.H. Otu, K. Sayood, Bioinformatics **19**(16), 2122–2130 (2003)
50. K. Carr, E. Murray, E. Armah, R.L. He, S.S.T. Yau, PLoS One **5**(3), e9550 (2010)
51. Y. Bakis, H.H. Otu, N. Tasci, C. Meydan, S. Yüzbaşıoğlu, O.U. Sezerman, BMC Bioinf. **14**(1), 20 (2013)

52. H. Minkowski, *Geometrie der Zahlen* (Chelsea, London, 1953)
53. R. Schwarz, M. Dayhoff, *Atlas of Protein Sequences and Structure* (National Biomedical Research Foundation, Washington, DC, 1978)
54. D.T. Jones, W.R. Taylor, J.M. Thornton, *Comput. Appl. Biosci.* **8**, 275–282 (1992)
55. P. Duhamel, H. Hollmann, *Electron. Lett.* **20**(1), 14–16 (1984)
56. F. Murtagh, *Comput. Stat. Q.* **1**, 101–113 (1984)
57. K. Tamura, D. Peterson, N. Peterson, G. Stecher, N. Nei, S. Kumar, *Mol. Biol. Evol.* **28**, 2731–9 (2011)
58. R.C. Edgar, *Nucl. Acids Res.* **32**(5), 1792–1797 (2004)
59. National Center for Biotechnology Information (NCBI) database. Available: <http://ncbi.nlm.nih.gov/protein>. Accessed 15 June 2013.
60. Global Initiative on Sharing All Influenza Data (GISAID) database. Available: <http://platform.gisaid.org>. Accessed 15 Jan 2013.
61. W.J. Murphy, E. Eizirik, S.J. O'Brien, O. Madsen, M. Scally, C.J. Douady, E. Teeling, O.A. Ryder, M.J. Stanhope, W.W. de Jong, M.S. Springer, *Science* **294**(5550), 2348–2351 (2001)
62. W.J. Murphy, E. Eizirik, W.E. Johnson, Y.P. Zhang, O.A. Ryder, S.J. O'Brien, *Nature* **409**(6820), 614–618 (2001)
63. V. Ranwez, F. Delsuc, S. Ranwez, K. Belkhir, M.K. Tilak, E.J. Douzery, *BMC Evol. Biol.* **7**(1), 241 (2007)
64. World Health Organization (WHO) website. Cumulative number of confirmed human cases of avian influenza A(H5N1) reported to WHO. Available: http://www.who.int/influenza/human_animal_interface/H5N1_cumulative_table_archives/en/index.html. Accessed Mar 2013.