Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 352 (2007) 177–180
BBRC
www.elsevier.com/locate/ybbrc

Influenza A H5N1 hemagglutinin cleavable signal sequence substitutions

Joel K. Weltman [a,*], Gail Skowron [a,b], George B. Loriot [c]

[a] Department of Medicine, Brown University School of Medicine, 4 Wildacre lane, Barrington, RI 02806-2630, USA
[b] Division of Infectious Diseases, Roger Williams Medical Center, Boston University School of Medicine and Brown University School of Medicine, Providence, RI 02912, USA
[c] Center for Computation and Visualization, Brown University, Providence, RI 02912, USA

Received 26 October 2006
Available online 10 November 2006

Abstract

Eleven influenza A H5N1 hemagglutinin N-terminal cleavable signal sequences, coded by single nucleotide substitutions relative to reference A/Viet Nam/1203/2004, were identified by BLASTN search of GenBank and were characterized by molecular modeling. The signal sequences statistically segregated into two classes of states. Members of one class were uncharged and conformationally compact while members of the second class each carried a 2+ electric charge and were conformationally extended. Virtual signal sequences, not found on GenBank and based upon hypothetical transversions in the third codon, had molecular characteristics intermediate to those of the two classes of actual signal sequences. The high incidence of non-synonymous substitutions (63.6%), the high transition/transversion ratio (10/1) and the results of molecular modeling all suggest that the N-terminal cleavable signal sequence is mutationally evolving more rapidly than proteins which must assume specific conformational states in the mature influenza virion.

Keywords: Influenza; H5N1; Avian; Hemagglutinin; Signal sequence; Mutations; Molecular mechanics; Evolution; Transitions; Transversions; Conformation

An analysis of single nucleotide substitutions in the influenza H5N1 hemagglutinin signal sequence based upon bioinformatics of the viral nucleic acid gene and molecular mechanics of the translated viral gene products is presented here. The cleavable, N-terminal 16-mer signal sequence of the hemagglutinin is essential for productive infection by the influenza A virus [1] and therefore, it is important to understand the fundamental principles that govern the structure, function, and evolution of the influenza hemagglutinin signal sequence.

Materials and methods

The signal sequence of the lethal, human isolate H5N1 influenza viral strain A/Viet Nam/1203/2004, was used as the reference for a BLASTN [2] search of the GenBank non-redundant database for influenza H5N1 strains with a hemagglutinin cleavable signal differing from that of the reference strain by a single nucleotide substitution. Signal sequences reported up to and including 2004, the year of isolation of the reference strain, were included in this study. The DNA sense representation of the antisense single stranded RNA signal sequence coding for the reference A/Viet Nam/1203/2004 viral signal sequence is: **atggagaaaatagtgcttcttttttgcaatagtcagtcttgttaaaagt**. The amino acid sequence of the reference hemagglutinin signal sequence is: **MEKIVLLFAIVSLVKS**.

Molecular models of each of the hemagglutinin signal sequences were constructed with TINKER [3] using the amoebapro force field [4]. The hemagglutinin signal sequences were modeled with the carboxy termini *N*-methyl capped so as to mimic the covalent structure of the hemagglutinin precursor. The geometry of each 16-mer signal sequence peptide was optimized using TINKER MINIMIZE.EXE. Each geometry optimized signal sequence peptide was characterized with TINKER ANALYZE.EXE to obtain the total potential energy, the electric charge, the dipole moment and the radius of gyration. Non-parametric Mann–Whitney *U* statistical tests of significance [5] of the data were performed with MATLAB (Version 7.0.1.24704). Descriptive statistics, i.e., median, mode, range, mean, and skewness, were obtained with MS-Excel.

* Corresponding author. Fax: +1 401 863 9265.
E-mail address: joel_weltman@brown.edu (J.K. Weltman).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.10.184

Results and discussion

Identification of influenza A H5N1 hemagglutinin signal sequences

Eleven influenza A H5N1 hemagglutinin signal sequences with single nucleotide substitutions relative to the reference sequence were identified by the BLASTN search and are listed in Table 1 along with their nucleotide substitution and amino acid designation. Four synonymous substitutions (36.4%) and seven non-synonymous substitutions (63.6%) were reported by the BLASTN search. Ten single nucleotide substitutions in the signal sequence are transitions relative to the reference signal sequence. The six transitions in signal sequences K3E, K3R, K3K, A9A, K15E, and S16G involve purine bases while the four transitions in signal sequences F8L, F8F, V11A, and V14V involve pyrimidines. The purine transition that occurred at the third position of codon 3 was synonymous for LYS. The pyrimidine transition at the third position of codon 8 was synonymous for PHE. Synonymous transitions also occurred at the third positions of codons 9 and 14. One of the 11 nucleotide substitutions shown in Table 1, at the first nucleotide of the sixth codon, is a cytosine ⇒ adenine transversion that causes a conservative, non-synonymous LEU ⇒ ILE substitution at the sixth amino acid position of the signal sequence (designated L6I).

Characterization of influenza A H5N1 hemagglutinin signal sequences

Molecular characterizations of the influenza A H5N1 hemagglutinin signal sequences that were identified by BLASTN search are given in Table 1. The median potential energy, electric charge, dipole moment, and radius of gyration of the hemagglutinin signal sequences shown in Table 1 were −269.1877 kcal/mol, +2 electron units, 41.249 D, and 14.951 Å, respectively. In all cases, the mode equaled the median. The mean potential energy, electric charge, dipole moment, and radius of gyration of the hemagglutinin signal sequences shown in Table 1 were −296.6158917 kcal/mol, +1.666666667 electron units, 36.52133 D, and 14.01383333 Å, respectively. In all cases, the parameter mean was less than the median, indicating non-normal distributions of molecular parameters. Accordingly, the hemagglutinin signal sequences were divided into two classes. Class 1 was defined as signal sequences K3E and K15E. Class 2 was defined as all of the other 10 signal sequences. The null hypothesis for the statistical identity of signal sequence classes 1 and 2 was rejected by the Mann–Whitney U test for all four molecular parameters with $p = 0.0303$ for each of the four. The probability that classes 1 and 2 do not represent statistically distinguishable molecular states is equal to 0.0303^4, or 8.4×10^{-7}. No other distribution of the signal sequences permitted such a definitive identification of two

Table 1
Influenza A H5N1 hemagglutinin signal sequence characterizations

Virus strain	GenBank Accession No.	Reference codon	Substituted codon	Signal sequence designation	Total potential energy (kcal/mol)	Total electric charge (electron units)	Dipole moment (D)	Radius of gyration (Å)
A/Viet Nam/ 1203/2004	AY651334			Reference	−269.1877	2.00	41.249	14.951
A/chicken/Viet Nam/159/2004	ABE97618	aaa	gaa	K3E	−425.2869	0.00	22.791	8.487
A/chicken/Viet Nam/c58/2004	AAW80718	aaa	aga	K3R	−299.5118	2.00	25.196	15.060
A/duck/Viet Nam/40/2004	DQ497673	aaa	aag	K3K	−269.1877	2.00	41.249	14.951
A/chicken/Sarburi/Thailand/ CU-27/2004	AAZ29968	ctt	att	L6I	−267.0616	2.00	43.931	14.825
A/chicken/Chachoengsao/ Thailand/ CU-11/2004	AAZ29955	ttt	ctt	F8L	−258.0824	2.00	43.457	15.113
A/goose/Hong Kong/739.2/ 2002	AY575871	ttt	ttc	F8F	−269.1877	2.00	41.249	14.951
A/chicken/Viet Nam/39/2004	AY651342	gca	gcg	A9A	−269.1877	2.00	41.249	14.951
A/mynas/Ranong/Thailand/ CU-209/2004	AAZ29981	gtc	gcc	V11A	−277.2001	2.00	30.006	16.382
A/duck/Hong Kong/821/2002	AAV97601	gtt	gtc	V14V	−269.1877	2.00	41.249	14.951
A/chicken/Suphanburi/ Thailand/ CU-9/2004	DQ083557	aaa	gaa	K15E	−418.1893	0.00	21.969	8.562
A/chicken/Hong Kong/NT93/ 2003	AAT73294	agt	ggt	S16G	−268.1201	2.00	44.661	14.982

Codons are listed in the DNA sense representation. Amino acid numbering is relative to the N-terminus.