INFLUENZA DIFFERENTIATION AND EVOLUTION*

Krzysztof Bartoszek

Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden
krzbar@chalmers.se

Pietro Liò†

Computer Laboratory, University of Cambridge
Cambridge, United Kingdom
pl219@cam.ac.uk

Anil Sorathiya

Computer Laboratory, University of Cambridge
Cambridge, United Kingdom
as833@cam.ac.uk

(Received February 18, 2010)

The aim of the study is to do a very wide analysis of HA, NA and M influenza gene segments to find short nucleotide regions, which differentiate between strains (*i.e.* H1, H2, ... *etc.*), hosts, geographic regions, time when sequence was found and combination of time and region using a simple methodology. Finding regions differentiating between strains has as its goal the construction of a Luminex microarray which will allow quick and efficient strain recognition. Discovery for the other splitting factors could shed light on structures significant for host specificity and on the history of influenza evolution. A large number of places in the HA, NA and M gene segments were found that can differentiate between hosts, regions, time and combination of time and region. Also very good differentiation between different Hx strains can be seen. We link one of our findings to a proposed stochastic model of creation of viral phylogenetic trees.

PACS numbers: 87.23.Kg, 87.10.Vg, 87.18.Tt

* Presented at the Summer Solstice 2009 International Conference on Discrete Models of Complex Systems, Gdańsk, Poland, June 22–24, 2009.

† Corresponding author.

1. Introduction

Statistical studies have provided interesting insights into the past influenza epidemics and pandemic. Every year the spreading of seasonal influenza causes significant mortality, and it cost billions of euros in health care expenses and with the risk of hospital-acquired infections [1–3]. Should the strains of influenza mutate to a strain that can quickly pass between birds and humans and then humans–humans it could cause a pandemic as it could spread to different parts of the world through human travel, migratory birds and in each place rapidly between humans.

The influenza virus is a common cause of respiratory infection all over the world. It infects not only humans, but also other species including avian and swine. Influenza A subtype (H1N1) could cause the next pandemic, if a new influenza A subtype has the ability to spread between humans efficiently. We have performed bioinformatic analysis to investigate the evolution of the HA, M and NA gene among different species. It would be important to know specific combination of viral RNA segments, hosts, regions, time and combination of time and region which have significant for host specificity and on the history of influenza evolution. Here we present a statistical analysis of influenza, and discuss how these statistics can provide insights on structures significant for host specificity and on the history of influenza evolution.

2. Data

Sequences of influenza strains were downloaded from GISAID [4]. Sequences from the HA, M and NA gene segments were downloaded. Each gene segment was considered separately. As some HA sequences were extremely short compared to the others these were removed. Each gene segment was downloaded with related information to it (*e.g.* its time period, geographical region and host). Unfortunately, for the sequences from the HA gene segment we did not have any information other than their strains. A summary of the downloaded gene segment sequences is presented in Table I.

In Table I the numbers of sequences do not always add up, there are two reasons for this. One is that for some sequences there is missing data *e.g.* host information but this was negligible. The main reason was that there were also sequences from the Australia and Oceania region and these were not considered in the regional analysis. The motivation was that there were relatively few of them compared to the other regions and when they were included nothing interesting was seen for them. Also a word is needed to motivate the method of regional splitting and temporal splitting. The temporal splitting was done in a fashion to have a equal balance between time bins. Unfortunately, the further we go back in time the fewer and fewer sequences we have. This type of binning has the consequence of ignoring

TABLE I

Breakdown of downloaded sequences. Abbreviations are: EA — Eastern Asia, CA — Central Asia, NA — North America, AO — Australia and Oceania, MEI — Middle East,Indian Peninsula, WE — Western Europe, SA — South America, EES — Eastern Europe Scandinavia, UKII — UK, Ireland, Iceland, A — Africa, EuA — Eurasia, As — Americas, Av — Avian, Sw — Swine, H — Human, Eq — Equine, Env — Environment, Om — Other mammals, Un — Unknown (these abbreviations will be used in subsequent tables).

HA gene segment

6052 sequences — alignment length: 2063 bases

H1	H2	H3	H4	H5	H6	H7	H8
904	132	1608	158	1833	286	509	17

H9	H10	H11	H12	H13	H14	H15	H16
387	65	71	26	27	4	7	15

HA gene segment H1N1

1851 sequences — alignment length: 2063 bases

EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
217	51	980	192	23	64	83	96	107	22

Human	Swine	Avian	Lab
1570	191	77	10

M gene segment

6775 sequences — alignment length: 1101 bases

EuA+Africa	As	1902–1999	2000–2004	2005–2009
2670	3243	2051	2227	2494

EuA+Africa 1902–1999	EuA+Africa 2000–2004	EuA+Africa 2005–2009
616	956	1053

As 1902–1999	As 2000–2004	As 2005–2009
1346	650	1247

Av	Sw	H	Eq	Env	Om	Un
2578	320	3650	90	101	18	18

NA gene segment

9661 sequences — alignment length: 1741 bases

EuA	As	1902–1999	2000–2004	2005–2009
4475	4160	2433	3244	3984

EuA 1902–1999	EuA 2000–2004	EuA 2005–2009
815	1729	1931

As 1902–1999	As 2000–2004	As 2005–2009
1502	842	1816

Av	Sw	H	Eq	Env	Om	Un
3827	461	5121	89	102	30	19

any changes that happened inside the time bins. The main motivation for geographic splitting was land and water masses. The whole of Eurasia (and effectively with Africa) is connected by land while it is separated from the Americas (the South and North also connected by land) and Australia and Oceania by water. These divisions of time and space are very rough and are such to keep the number of bins at a low level.

3. Method

The method employed to analyze a set of sequences is a very preliminary one. The aim is to have a first look at the data and see whether anything new can be discovered without explicit mathematical modelling. For computational reasons mathematical models (*e.g.* Markov models) of alignment data nearly always assume independence of columns. We are not aware of any satisfactory and computationally effective model that considers even short dependencies between columns in such a large alignment. Therefore we want to see how much can be discovered without the additional knowledge a model would bring. In this paper we will introduce the following terminology, a metasubsequence of length k will be vector of k consecutive positions in the alignment and by the value of the metasubsequence we will mean the actual values (for a given sequence, A, C, G, T in a nucleotide alignment) that are observed in the respective columns of the alignment.

3.1. Alignment

The alignment of the sequences was done in `Clustal 2.0.10` [5]. Due to the large number of sequence it was done with the option approximate and all others default. Prior to the alignment sequences which were considerably shorter than the others had to be removed. In the case of the H1N1 sequences this was roughly 1500 sequences. Had this not been done the contingency table test did not look at splits between hosts, regions, *etc.* but between short and long sequences.

3.2. Contingency table

The contingency table test (or chi-square test) is described in detail in *e.g.* [6] but we will make a short overview of it here. Let us assume we have N objects in a two-way categorization table with an arbitrary number of rows and columns. In our case the rows will be the different possible values of metasubsequences the regions take and the columns will describe the different possible categories the sequences can take). Such a table can look like the one in [6].

					Total
	Y_{11}	Y_{12}	...	Y_{1k}	$y_{1\cdot}$
	Y_{21}	Y_{22}	...	Y_{2k}	$y_{2\cdot}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	Y_{n1}	Y_{n2}	...	Y_{nk}	$y_{n\cdot}$
Total	$y_{\cdot 1}$	$y_{\cdot 2}$	...	$y_{\cdot k}$	y

Under the null hypothesis that there is no association between row and column categories each entry Y_{jk} will be a random variable with mean value of $E_{jk} = \frac{y_{j\cdot} y_{k\cdot}}{y}$. The test statistic is $\sum_{jk} \frac{(Y_{jk}-E_{jk})^2}{E_{jk}}$ and under the null hypothesis has a chi-square with $(r-1)(c-1)$ degrees of freedom distribution. An important assumption of this test is that the observations leading up to the counts are independent. In our case they are not due to the phylogenetic history. Taking this into account is a topic for further study and due to this we cannot assign any formal statistical significance to the results. This problem of dependence has been very recently looked in a simulation study in [7] however the authors consider alignment columns independently. The procedure of working with the contingency table is described in the following algorithm

1: **for** metasubsequence length $i = 0$ to n **do**
2: **for** column $j = 1$ to length of alignment **do**
3: build contingency table for metasubsequence at position j to $j + i$ $\{i$ is from $0\}$
4: calculate p-value of contingency table
5: **end for**
6: **end for**
7: return those contingency tables with their positions and metasubsequences that are below some cut-off p-value
8: go through the returned contingency tables and see whether they are interesting.

The cut-off p-value has to be extremely low (it was taken to be 1^{-183}) due to the fact that we have at each position very often one sequence which is totally different from the others and such a sequence could immediately generate a significant table. Unfortunately very often we have a huge number of tables, in fact in some cases every single consecutive metasubsequence of the alignment was significant due to this effect. Our choice of scoring is presented in the next section. Up until now most analysis have been treating positions independently *e.g.* [8] (nucleotide alignment) [9] (protein alignment). An independent site analysis by a direct application of entropy [9–11] methods

is straightforward, quick and the results are immediate and do not require any postprocessing. The problem is that it might miss columns that are significant only when a combination of them is considered. Here we make an attempt to do a large scale sequence analysis to find metasubsequences of up to 25 bases which categorize the sequences. We are especially interested in metasubsequences of up to 25 bases because this is what we expect that the length of the probe on the Luminex microarray will be.

3.3. Scoring contingency tables

The difficulty with the contingency table is that if the alignment is very "noisy" (either due to misalignment or a large number of sequences that have mutations) nearly all positions in the alignment will turn significant. Therefore some other method than the p-value of scoring each contingency table has to be used. We adopted the following strategy for each contingency table,

1: remove all rows that have in total less than $p\%$ of the sequences
2: remove all columns that have in total less than $p\%$ of the sequences
3: $N =$ number of rows
4: **for** each column i **do**
5:

$$\text{Score}_i = 1 - \frac{-\sum_{\text{each row j}} p_{ij} \log p_{ij}}{\log N}$$

6: **end for**
7: **return** $\max_i$ Score_i.

We look at each column of the contingency table (which represents some grouping of the sequences) and see what is the estimated entropy of the sequences in this group $-\sum_{\text{each row j}} p_{ij} \log p_{ij}$ dividing by $\log N$ the maximal entropy possible. The score will be 1 if we can predict the value of the metasubsequence perfectly, *i.e.* there is exactly one value of a given metasubsequence in the given group and 0 if we cannot say anything. To get the score for a given position (in Figs. 2, 3, 4 and 5) we take the maximum over all metasubsequences of length $1 \ldots 25$. In our analysis we decided to ignore gaps in the HA and M gene segment alignments as these were aligned with very few gaps (except on the edges) while in the analysis of the NA gene segment gaps were treated as a fifth residue (here there was a large number of gaps inside the alignment). The methodology presented is naturally very heuristic due to the large computational intensity and complexity of the problem. However, it guarantees that if a given metasubsequence is present in enough sequences and its values split perfectly between groups then the method will find it.

3.4. Implementation

The described method was implemented as a `Perl` [12] script. For working with the alignments it uses `BioPerl` [13] and for calculating the p-value of the χ^2 statistic the module `Statistics::Distributions` [14] was used.

3.5. Phylogenetic tree

HA gene segment of 6052 sequences was aligned using `Clustal 2.0.10` tool [5]. This alignment of 2063 bases length from H1 to H16 viral strains was used to generate the phylogenetic tree which describing the evolution of strains from one to another (see figure 1). Colour coding of the parenthesis in Fig. 1 represents strains from H1 to H16 and `Dendroscope 2.4` tool [15] was used to visualize the phylogenetic tree.

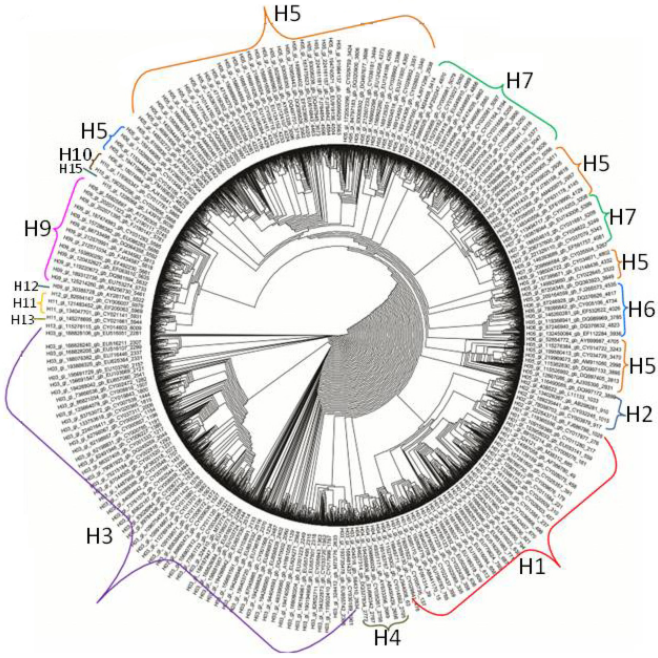

Fig. 1. Phylogeny tree of influenza strains; Each colour represents each strain of influenza virus.

4. Results

The biological questions posed in the analysis are the following (each item is done on a different sequence set),

- Can we find regions of length 20–25 bases the distinguish between Hx strains (for the Luminex microarray)?

- Can we find short regions up to 25 bases which will distinguish between hosts and regions in H1N1 sequences?
- Can we find short regions up to 25 bases which will distinguish between hosts, regions, time periods in M sequences, can any temporal and geographical interactions be found?
- Can we find short regions up to 25 bases which will distinguish between hosts, regions, time periods in NA sequences, can any temporal and geographical interactions be found?

In all cases interesting places were found.

4.1. Hx analysis

Only H1, H2, H3, H4, H5, H6 and H7 sequences were considered. The other ones were too few in numbers compared to these, so were discarded in the analysis as they only acted as "noise". All in all 292 places of length 20–25 bases were found in the HA genomes that split (nearly perfectly) between the strains. Some of them built up together to form a larger region so we could actually distinguish about 6 regions in the genome where differences between the strains can be found. It is not possible to present all the results but a nice example contingency table can be shown, Table II. The values inside the table are the fraction of the number of strains having the given value of the metasubsequence. The fact that for a strain a column does not add up to 1 is due to it having less than 100 sequences in some rows and for clarity of the presentation if a row contained less than 100 sequences it was not written out. The scoring function failed to cut-down on the number of positions as all of the splits were perfect. In Fig. 2 we can see how that the scores were nearly 1 all along the gene segment.

TABLE II

Example contingency table splitting Hx strains of HA gene segment, 292 such places were found. We can see that combinations of different positions are needed to differentiate between all.

Position 1687	H1	H2	H3	H4	H5	H6	H7
TGGGAC**T**TATGAC**C**A**TGATGTA**TA**C**	0	0	0.07	0	0	0	0
CAACAC**T**TATGAC**C**A**TACTCAA**TA**C**	0	0	0	0	0	0	0.24
CAACAC**G**TATGAC**C**A**TACTCAA**TA**C**	0	0	0	0	0	0	0.21
TGGAAC**T**TATGAC**C**A**TGATGTA**TA**C**	0	0	0.53	0	0	0	0
TGGAAC**T**TATGAC**C**A**CGATGTA**TA**C**	0	0	0.1	0	0	0	0
TGGAAC**T**TATGAC**T**A**TCCAAAA**TA**T**	0.63	0	0	0	0	0	0
CGGAAC**G**TATGAC**T**A**CCCGCAG**TA**T**	0	0	0	0	0.69	0	0

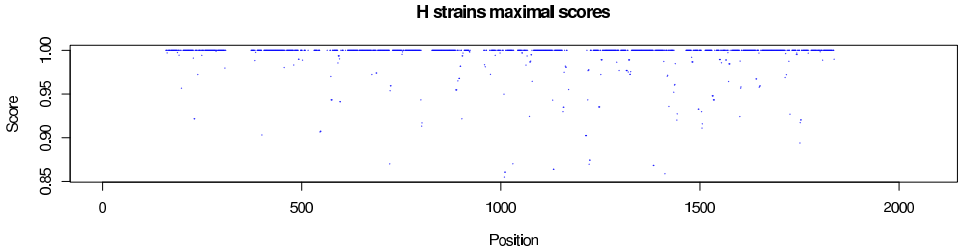

Fig. 2. Scores along the alignment of the HA gene segment. We can see that nearly all along the gene segment we have a score of 1. The visible gaps in the graph are due to gaps in the alignment in those places.

4.2. H1N1 analysis

Two types of analysis were done, whether any place specific for hosts or region can be found.

In the host analysis (Table III) five places specific for human hosts were found. They were at positions 282–307, 1085–1110, 1133–1158, 1219–1244 and 1570–1575 in the alignment. About 83% of all human had the same sequence.

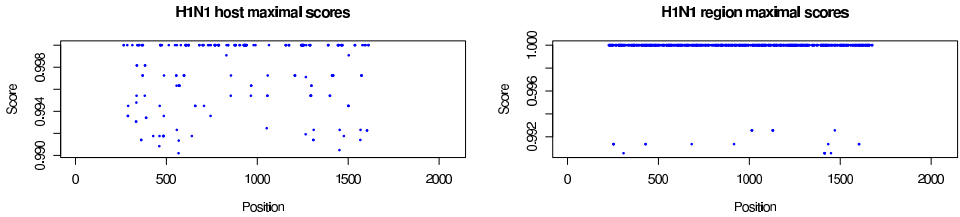

Fig. 3. Top: scores along the alignment of the HA H1N1 gene segment for splits between hosts. All along the gene segment we have positions achieving high scores (cut-off for graph is 0.99). A few of the interesting examples are in Table III. Bottom: scores along the alignment of the HA H1N1 gene segment for splits between regions. All along the gene segment we have positions achieving high scores (cut-off for graph is 0.99). A few of the interesting examples are in Table IV.

As previously the columns do not add up to 1 due to many rows having very small numbers of sequences and we can see changes specific for all three hosts. When the analysis was done according to regions the North American region stood out significantly (Table IV) in the HA gene segment.

TABLE III

Examples of top scoring (above 0.95) host specific changes HA genome of H1N1, the score is in brackets next to the position number. There were in total 70 such places found.

Position 286 (0.96)	Human	Swine	Avian
TGGGAAACCCAGAATGTGAATTACT	0.003	0.387	0.013
TGGGAAATCCAGAGTGTGAATCACT	0.042	0.021	0
TAGGAAACCCAGAATGCGAATCACT	0.1	0.01	0
TAGGAAACCCAGAATGCGAATTACT	0.734	0	0
TAGGAAACCCAGAATGCGAATTGCT	0.045	0	0

Position 396(0.96)	Human	Swine	Avian
A	0.004	0.728	0.065
G	0.996	0.272	0.935

Position 513(1)	Human	Swine	Avian
T	0.836	0.01	0
A	0.161	0.99	1

Position 579 (1)	Human	Swine	Avian
A	0.175	1	1
G	0.824	0	0

Position 1052(0.98)	Human	Swine	Avian
A	0.305	0.984	0.961
G	0.692	0	0

Position 1128 (0.96)	Human	Swine	Avian
GGTTTTATTGAGGG	0.002	0.408	0.052
GGATTCATTGAAGG	0.001	0.183	0.3
GGTTTCATTGAAGG	0.98	0.026	0.13

Position 1155 (0.98)	Human	Swine	Avian
ATGGTAGATGG	0.862	0	0
ATGATTGATGG	0.031	0.01	0
ATGATAGATGG	0.096	0.942	0.974

Position 1261 (0.97)	Human	Swine	Avian
TGAACTCTATAATCGAGAAAATGAA	0.03	0	0
TGAATTCTGTGATTGAGAAAATGAA	0.03	0	0
TGAACTCTGTTATCGAGAAAATGAA	0.024	0	0
TAAATTCTGTTATTGAAAAGATGAA	0.042	0.513	0
TGAATTCTGTAATCGAGAAAATGAA	0.036	0.01	0
TGAATTCTGTAATTGAGAAAATGAA	0.682	0	0.013
TGAATTCTGTAATTGAGAAGATGAA	0.045	0	0
TGAACTCTGTAATTGAGAAGATGAA	0.055	0	0

Position 1566 (0.99)	Human	Swine	Avian
AATGG	1	1	0.182
AACGG	0	0	0.818

Position 1582 (0.96)	Human	Swine	Avian
ACCCAAAATACTCAGAGGAAGCAAA	0.042	0.021	0
ATCCAAAATATTCCGAAGAATCAAA	0.825	0	0.013
ATCCAAAATATTCAGAGGAATCAAA	0.024	0	0
ATCCAAAATATTCAGAAGAATCAAA	0.027	0	0
ACCCAAAGTACTCAGAAGAATCAAA	0	0.225	0

TABLE IV

Examples of top scoring (above 0.95) regional differentiation of HA H1N1 sequences. There were 229 such positions found. Score is in brackets next to position number.

Position 286 (1)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
TGGGCAA	0	0	0.046	0	0	0.016	0	0	0.009	0
TTGGCAA	0.014	0.098	0.01	0.01	0	0.406	0	0.01	0.075	0
TGGGAAA	0.073	0.137	0.203	0.016	0	0.031	0	0.01	0.009	0
TAGGAAA	0.899	0.765	0.732	0.969	1	0.469	1	0.979	0.879	1
Position 333 (0.99)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
TATATAA	0.005	0	0.046	0.01	0	0.078	0	0	0.009	0
TACATAA	0.009	0.098	0.014	0	0	0.4212	0	0.01	0.075	0
TACATTG	0.972	0.882	0.909	0.984	0.913	0.5	0.988	0.99	0.888	1
Position 358 (1)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
AAAATGG	0.037	0.078	0.01	0.063	0	0.5	0	0.052	0.028	0
ACAATGG	0.092	0.059	0.188	0.016	0	0.031	0	0.01	0.019	0
AGAATGG	0.843	0.784	0.736	0.911	1	0.453	1	0.927	0.869	1
AGAACGG	0	0	0.045	0	0	0	0	0	0	0
Position 507(1)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
GGTGT	0.106	0.137	0.251	0.026	0	0.234	0	0.01	0.028	0
GGAGT	0.876	0.784	0.733	0.974	1	0.469	1	0.969	0.869	1
Position 649 (0.96)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
TATGGGG	0.982	0.863	0.573	0.922	0.826	0.531	0.843	0.979	0.869	1
TGTGGGG	0.005	0.039	0.363	0.068	0.174	0.016	0.157	0.01	0.019	0
TCTGGGG	0	0	0.059	0	0	0.328	0	0.01	0.084	0

Position 706 (1)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
AA	0.866	0.804	0.742	0.974	0.957	0.469	1	0.979	0.869	1
CT	0.009	0.039	0.044	0	0	0.359	0	0	0.084	0
CA	0.101	0.078	0.194	0.016	0	0.031	0	0.01	0.047	0
Position 902 (1)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
GGC	0.88	0.88	0.75	0.98	1.00	0.92	1	0.99	0.94	0.91
AAT	0	0	0.064	0.016	0	0.031	0	0.01	0	0
GGG	0.0046	0	0.0561	0	0	0.0312	0	0	0.028	0
GGT	0.074	0.059	0.120	0	0	0	0	0	0.028	0
Position 947 (1)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
TGCAC	0.0046	0.0588	0	0	0	0.1250	0	0.0104	0	0
TGCGA	0	0	0.04	0	0	0	0	0	0	0
TGCAA	0.018	0	0.069	0.016	0	0.031	0	0.01	0.028	0
TGTAC	0	0	0	0	0	0.125	0	0	0.084	0
TGTGA	0.88	0.8	0.77	0.98	1	0.72	1	0.98	0.87	1
TGTAA	0.092	0.137	0.124	0	0	0	0	0	0.019	0
Position 1422 (1)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
AGGACT	0.866	0.824	0.761	0.917	0.957	0.797	1	0.958	0.916	1
AGAACT	0.115	0.137	0.229	0.078	0.044	0.047	0	0.042	0.019	0
Position 1566 (0.95)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
AATGGAACTTA	0.899	0.765	0.688	0.974	1	0.453	0.988	0.979	0.907	0.909
AATGGCACATA	0.009	0.098	0.001	0	0	0.422	0	0	0.075	0
AATGGGACTTA	0.0783	0.137	0.248	0.0156	0	0.0313	0.0121	0.0104	0.009	0
AACGGCACATA	0.005	0	0.054	0	0	0.078	0	0	0	0
Position 1611 (1)	EA	CA	NA	AO	MEI	WE	SA	EES	UKII	A
AG	0.419	0.451	0.011	0.182	0.044	0.031	0.048	0.073	0	0
AA	0.581	0.529	0.986	0.818	0.957	0.969	0.952	0.927	1	1

4.3. M results

The M gene segment sequences were analyzed to find differences between hosts, regions, time periods and geographical and temporal interactions. The results are presented in Fig. 4 and Tables V, VI, VII and VIII.

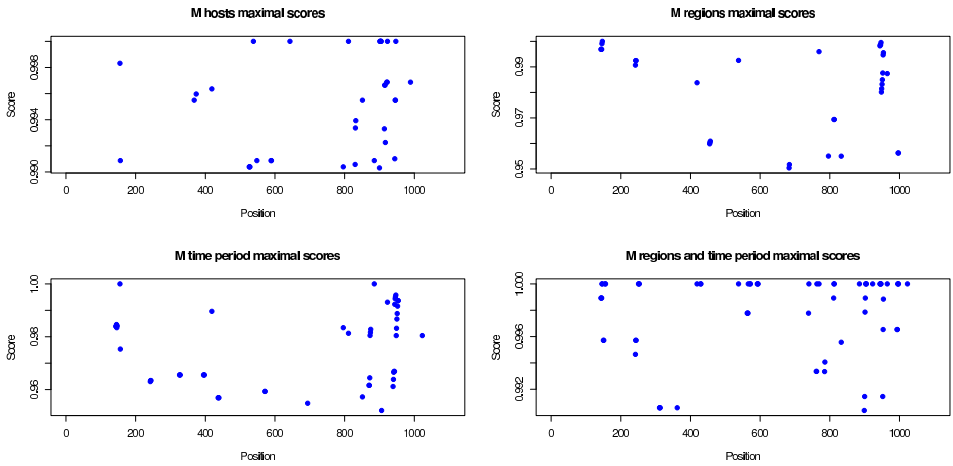

Fig. 4. Top left: scores along the alignment of the M gene segment for splits between hosts, cut-off for graph is 0.99. Top scoring examples are in Table VII. Top right: scores along the alignment of the M gene segment for splits between regions, cut-off for graph is 0.95. Top scoring examples are in Table V. Bottom left: scores along the alignment of the M gene segment for splits between time periods, cut-off for graph is 0.95. Top scoring examples are in Table VI. Bottom right: scores along the alignment of the M gene segment for splits between regions combined with time periods, cut-off for graph is 0.99. Top scoring examples are in Table VIII.

TABLE V

Top scoring (above 0.95) region specific changes of M gene segment.

Position 143 Score 0.99	Eurasia & Africa	Americas	Position 242 Score 0.98	Eurasia & Africa	Americas
CAGAGACT	0.709	0.994	ATTTT	0.867	0.985
CAGAAACT	0.278	0	ATGTT	0.067	0
Position 538 Score 0.99	Eurasia & Africa	Americas	Position 683 Score 0.95	Eurasia & Africa	Americas
TC	0.9	0.999	TTGC	0.627	0.99
GC	0.1	0.001	TCGC	0.369	0.005
Position 768 Score 0.96	Eurasia & Africa	Americas	Position 944 Score 0.99	Eurasia & Africa	Americas
CTTGAAAATTT	0.826	0.977	CCTTCTACGGCAGG	0.059	0
ATTGAAAATTT	0.149	0	CCTGCTACGGCAGG	0.071	0
			CCTTCTACGGAAGG	0.814	0.984

TABLE VI

Top scoring (above 0.95) time period specific changes of M gene segment.

Position 143 Score: 0.97	Until 1999	2000–2004	2005–2009	Position 242 Score: 0.96	Until 1999	2000–2004	2005–2009
CAGAGACT	0.987	0.920	0.761	ATTTT	0.973	0.972	0.861
CAGAAACT	0.001	0.072	0.232	ATGTT	0.0005	0.002	0.07
Position 326 Score: 0.97	Until 1999	2000–2004	2005–2009	Position 571 Score: 0.96	Until 1999	2000–2004	2005–2009
AATGG	0.893	0.991	0.996	AC	0.996	0.888	0.765
AACGG	0.106	0.009	0.0036	AT	0.004	0.106	0.235
Position 871 Score: 0.95	Until 1999	2000–2004	2005–2009	Position 939 Score: 0.95	Until 1999	2000–2004	2005–2009
GATATTG	0.982	0.899	0.966	GAGGCCCTTCTAC	0.001	0.146	0.233
GATACTG	0.003	0.093	0.0008	GAGGGCCTTCTAC	0.956	0.764	0.676
				GAGGGCCTGCTAC	0	0.042	0.031
Position 1023 Score: 0.98	Until 1999	2000–2004	2005–2009				
GTCAT	0.992	0.994	0.827				
ATCAT	0.001	0.002	0.164				

TABLE VII

Most interesting of the top scoring (above 0.95) host specific changes of M gene segment. Score in brackets after position number. There were 73 such positions found.

Position 410 (097)	Avian	Swine	Human	Equine	Position 478 (0.96)	Avian	Swine	Human	Equine	Position 914 (0.98)	Avian	Swine	Human	Equine
AT	0.003	0.091	0.907	0	AC	0.991	0.928	0.082	1	CGTCGC	0.98	0.85	0.097	0.8
GT	0.997	0.906	0.092	1	GC	0.0035	0.072	0.918	0	CGTATC	0.0004	0	0.22	0
										CGACTC	0	0.044	0.57	0
										CGATTC	0	0.003	0.039	0

Position 921 (0.98)	Avian	Swine	Human	Equine	Position 933 (0.96)	Avian	Swine	Human	Equine
TCAAACACGG	0	0.0406	0.523	0	TGAAA	0.989	0.875	0.520	0.944
TTAAACACGG	0.00233	0	0.25	0.011	TAAAA	0.006	0.009	0.207	0.056
TTGAACACGG	0	0.013	0.042	0	TTAAA	0	0.019	0.27	0
TCAGACACGG	0	0	0.057	0					
TTAAATACGG	0.976	0.872	0.095	0.9					

Position 939 (0.98)	Avian	Swine	Human	Equine	Position 1013 (0.95)	Avian	Swine	Human	Equine
GAGGGCCTT	0.888	0.947	0.724	0.911	GTTGACGAT	0.952	0.869	0.095	0.889
GAGGGCCTG	0.054	0	0.01	0	GCTGACGAT	0.0008	0.009375	0.281	0
GAGGCCCTT	0.009	0	0.243	0	GCTGACGAC	0	0.059	0.589	0

TABLE VIII

Top scoring (above 0.95) regions and time period specific changes of M gene segment. Score in brackets after position number. African sequences not included in Eurasia.

Position 98 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 143 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GTTCTCTCTAT	0.130	0.937	0.904	0.981	0.983	0.886	CAGAAACTTGAAGATGT	0.002	0.166	0.393	0	0	0
GTTCTTTCTAT	0.039	0.053	0.069	0.001	0.002	0.112	CAGAGACTTGAAGATGT	0.818	0.659	0.310	0.939	0.971	0.794
							CAGAAACTTGAGGATGT	0	0.002	0.12	0	0	0
							CAGAGACTGGAAAGTGT	0	0	0.038	0	0	0.111
							CAGAGACTTGAGGATGT	0.065	0.103	0.087	0.039	0.009	0.078
Position 185 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 200 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GAGGCACT	0.044	0.006	0.002	0.449	0.392	0.232	TGGCTAAAG	0.992	0.951	0.93	0.99	0.992	0.92
GAGGCTCT	0.920	0.983	0.979	0.482	0.591	0.711	TGGTTAAAG	0.005	0.015	0.006	0.006	0.008	0.073
Position 204 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 221 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TAAAGACAAGACCAAT	0.929	0.976	0.956	0.906	0.942	0.993	CTGTCACCTCT	0.963	0.955	0.839	0.975	0.991	0.886
TAAAGACAAGACCGAT	0.06	0.017	0.01	0.085	0.051	0.002	TTGTCACCTCT	0.018	0.038	0.058	0.018	0.005	0.111
Position 242 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 293 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
ATTTT	0.945	0.949	0.779	0.984	0.98	0.988	CGTAGACGGTTTGT	0.003	0.088	0.06	0.0033	0.005	0
ATGTT	0.002	0.005	0.13	0	0	0	CGTAGACGATTTGT	0.112	0.164	0.047	0.022	0.002	0
							CGTAGACGCTTTGT	0.864	0.728	0.86	0.952	0.951	0.978
Position 323 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 354 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GGAAATGG	0.06	0.346	0.578	0.015	0.04	0.006	CAGT	0.961	0.985	0.983	0.776	0.754	0.851
GGGAACGG	0.058	0.01	0.002	0.073	0.002	0.002	CGGT	0.015	0.003	0.008	0.215	0.24	0.149
GGGAATGG	0.846	0.621	0.418	0.871	0.938	0.99	AAACT	0.916	0.627	0.357	0.932	0.986	0.997
Position 359 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 368 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AAATT	0.024	0.113	0.061	0.014	0.011	0.001	A	0.974	0.973	0.937	0.997	0.903	0.671
AAGCT	0.057	0.256	0.573	0.053	0.003	0.002	C	0.026	0.027	0.062	0.002	0.097	0.329
Position 369 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 389 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.206	0.559	0.717	0.007	0.003	0.111	ACAT	0.969	0.74	0.782	0.994	0.875	0.629
G	0.791	0.441	0.283	0.993	0.997	0.889	ACGT	0.023	0.261	0.216	0.005	0.122	0.368

Position 392 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TTCCACGG	0.021	0.004	0.004	0.097	0.018	0.003
TTCCATGG	0.937	0.935	0.960	0.880	0.966	0.995

Position 416 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
CTCAG	0.946	0.925	0.922	0.978	0.946	0.779
CTAAG	0.01	0.048	0.065	0.001	0.003	0.111
CTCGG	0.002	0	0	0	0.003	0.103

Position 429 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
CCGGTGCACTCGC	0	0.002	0.126	0.002	0	0
CCGGTGCACTTGC	0.192	0.346	0.417	0.0215	0.019	0
CTGGTGCGCTTGC	0.084	0.236	0.098	0.006	0.005	0.004
CTGGTGCACTTGC	0.651	0.343	0.33	0.861	0.951	0.98

Position 469 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AG	0.904	0.782	0.931	0.997	0.971	0.985
CG	0.096	0.213	0.069	0.002	0.029	0.015

Position 479 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
CAGT	0.157	0.263	0.31	0.009	0.003	0.115
CGGT	0.237	0.444	0.46	0.006	0	0.002
CTGT	0.601	0.29	0.231	0.981	0.983	0.882

Position 484 (1)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
ACCAC	0.935	0.735	0.404	0.941	0.968	0.984
ACTAC	0.011	0.203	0.518	0.016	0.012	0.001

Position 496 (1)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GCATTTGG	0.146	0.289	0.215	0.678	0.978	0.729
GCTCTTGG	0.086	0.241	0.101	0	0	0
GCTTTTGG	0.294	0.434	0.649	0.024	0	0.112
GCCTTTGG	0.281	0.005	0.021	0.214	0.006	0.003
GCGTTTGG	0.002	0.001	0.004	0.039	0.003	0.067

Position 529 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
ATTGCCGA	0.029	0.025	0.062	0.005	0.106	0.399
ATTGCTGA	0.945	0.698	0.369	0.987	0.894	0.601
ATTGCAGA	0.01	0.244	0.538	0.001	0	0

Position 538 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TCACA	0.369	0.557	0.68	0.031	0	0.111
TCCCA	0.583	0.275	0.226	0.952	0.977	0.871
GCCCA	0.044	0.158	0.07	0.001	0.001	0

Position 563 (0.99)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TGGC	0.162	0.56	0.714	0.006	0	0.111
TGGT	0.826	0.434	0.285	0.987	0.998	0.887

Position 592 (1)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
CATGA	0.987	0.999	0.989	0.954	0.818	0.864
CACGA	0.002	0.003	0.051	0.016	0.008	0.119

Position 569 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
CTACCACCAA	0.279	0.292	0.197	0.005	0	0.111
CTATCACCAA	0.013	0.206	0.353	0	0	0
CTACTACCAA	0.033	0.066	0.04	0	0	0
CCATCACCAA	0	0.005	0.144	0	0	0
CAACAACCAA	0.573	0.288	0.225	0.95	0.932	0.868

Position 625 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GCTAA	0.982	0.993	0.941	0.937	0.96	0.873
GCAAA	0.002	0.003	0.051	0.016	0.008	0.119

Position 643 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
ATGGC	0.992	0.992	0.961	0.975	0.948	0.87
GTGGC	0.005	0.008	0.037	0.019	0.043	0.128

Position 683 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 689 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TTGC	0.162	0.55962	0.71415	0.006	0	0.111	A	0.008	0.242	0.591	0.01 0	.003	0.189
TCGC	0.056	0.35042	0.55081	0.006	0.006	0.002	G	0.992	0.758	0.409	0.988	0.995	0.811

Position 709 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 736 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
CATGC	0.073	0.051	0.066	0.014	0.003	0.111	CCTAG	0.154	0.745	0.448	0.923	0.983	0.99
CAGGC	0.903	0.938	0.929	0.924	0.992	0.889	CCTAA	0.016	0.227	0.528	0.016	0	0.005

Position 751 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 761 (0.99)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GGTTT	0.044	0.023	0.01	0.062	0.017	0.003	ATGA	0.984	0.734	0.448	0.964	0.995	0.966
GGTCT	0.948	0.923	0.967	0.928	0.974	0.994	ATAA	0.01	0.245	0.52	0.033	0	0

Position 769 (1)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 785 (0.99)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
CTTGAAAATTT	0.886	0.742	0.859	0.961	0.997	0.985	CTTA	0.045	0.127	0.094	0.014	0.005	0
ATTGAAAATTT	0.09	0.24	0.109	0.001	0	0	CATA	0.018	0.079	0.026	0.014	0.012	0.001
							CCTA	0.930	0.787	0.877	0.972	0.983	0.998

Position 796 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 811 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AG	0.002	0.01	0.002	0.002	0.018	0.127	CTGCA	0.005	0.031	0.033	0	0.018	0.121
CG	0.995	0.990	0.998	0.999	0.982	0.873	ATACA	0.05	0.0544	0.0427	0.007	0	0
							ATGCA	0.937	0.911	0.924	0.991	0.98	0.875

Position 796 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 811 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AG	0.002	0.01	0.002	0.002	0.018	0.127	CTGCA	0.005	0.031	0.033	0	0.018	0.121
CG	0.995	0.990	0.998	0.999	0.982	0.873	ATACA	0.05	0.0544	0.0427	0.007	0	0
							ATGCA	0.937	0.911	0.924	0.991	0.98	0.875

Position 820 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 833 (1)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TTCAAA	0.016	0.03	0.008	0.001	0.132	0.072	CT	0.987	0.902	0.896	0.955	0.997	0.998
TTCAAG	0.977	0.949	0.991	0.987	0.866	0.893	AT	0.002	0.098	0.103	0.003	0	0.002

Position 837 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 840 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TC	0.021	0.04	0.095	0.004	0.034	0.232	CTGC	0.023	0.008	0.013	0.085	0.002	0.001
TT	0.159	0.938	0.830	0.987	0.963	0.759	TTGC	0.948	0.964	0.945	0.890	0.974	0.982

Position 837 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 840 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TC	0.021	0.04	0.095	0.004	0.034	0.232	CTGC	0.023	0.008	0.013	0.085	0.002	0.001
TT	0.159	0.938	0.830	0.987	0.963	0.759	TTGC	0.948	0.964	0.945	0.890	0.974	0.982
Position 852 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 885 (1)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TAAT	0.028	0.021	0.062	0	0.092	0.328	TTGATCG	0.971	0.99	0.956	0.993	0.998	0.889
TCAT	0.942	0.934	0.895	0.966	0.903	0.67	CTGATCG	0	0	0.038	0.001	0	0.11
Position 899 (0.99)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 908 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TTCAAAT	0.961	0.973	0.907	0.993	0.902	0.671	AT	0.987	0.802	0.842	0.968	0.852	0.744
TCCAAAA	0.028	0.024	0.057	0	0.091	0.325	GT	0.002	0.184	0.151	0.027	0.148	0.255
Position 923 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 933 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AGA	0.01	0.019	0.001	0.006	0.134	0.038	TGAAA	0.104	0.797	0.774	0.986	0.772	0.402
GAA	0.119	0.001	0.002	0.053	0	0.006	TAAAA	0.028	0.025	0.062	0.007	0.1	0.328
AAA	0.867	0.978	0.997	0.928	0.862	0.954	TTAAA	0.052	0.177	0.149	0	0.115	0.256
Position 935 (0.95)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 965 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AAAAGAGGGCCTGCTACGGCAGG	0	0.097	0.07	0	0	0	GAGTC	0.974	0.983	0.972	0.953	0.972	0.868
AAAAGAGGGCCTTCTACGGCAGG	0	0.061	0.084	0	0	0	AAGTC	0.003	0.001	0	0.02	0.02	0.11868
AAAAGAGGCCCTTCTACGGAAGG	0	0.135	0.138	0.002	0.123	0.239							
AAAAGAGGGCCTTCTACGGAAGG	0.930	0.599	0.543	0.929	0.802	0.685							
Position 971 (0.96)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 980 (0.97)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
ATGAGGGAGGA	0.01	0.04	0.004	0.022	0.054	0.129	GAATATC	0.787	0.445	0.362	0.899	0.948	0.994
ATGAGGGAAGA	0.977	0.957	0.994	0.978	0.945	0.867	GAGTATC	0.195	0.371	0.279	0.025	0.046	0.003
							GAGTACC	0.006	0.177	0.355	0.011	0.002	0.001
Position 990 (0.98)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 993 (1)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AGAA	0.013	0.001	0.01	0.092	0.017	0.014	AGCAACAG	0.026	0.084	0.01	0.008	0	0
AGGA	0.974	0.994	0.985	0.903	0.969	0.984	AACAGCAG	0.938	0.889	0.953	0.942	0.977	0.998
Position 1023 (1)	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09							
GTCAT	0.997	0.992	0.972	0.99	0.995	0.692							
ATCAT	0	0.003	0.021	0.001	0.0015	0.297							

4.4. NA results

The NA gene segment sequences were analyzed to find differences between hosts, regions, time periods and geographical and temporal interactions. The results are presented in Fig. 5 and Tables IX, X, XI and XIII.

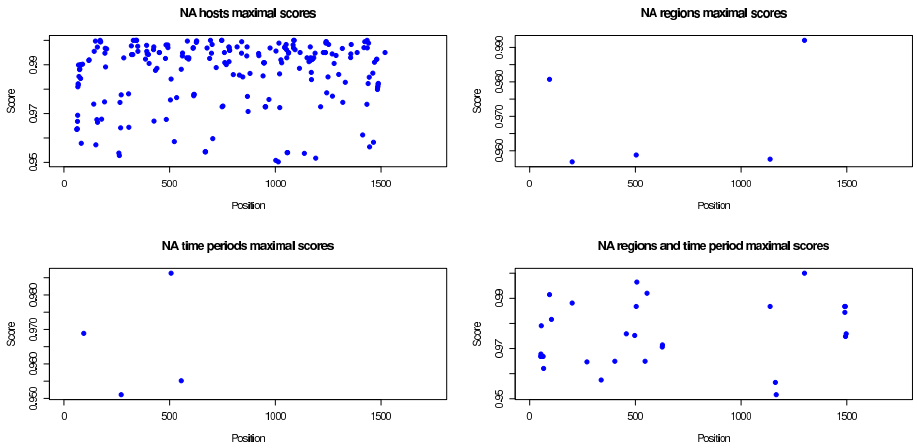

Fig. 5. Top left: scores along the alignment of the NA gene segment for splits between hosts, cut-off for graph is 0.95. Top scoring examples are in Table XII. Top right: scores along the alignment of the NA gene segment for splits between regions, cut-off for graph is 0.95. Top scoring examples are in Table X. Bottom left: scores along the alignment of the NA gene segment for splits between time periods, cut-off for graph is 0.95. Top scoring examples are in Table XI. Bottom right: scores along the alignment of the NA gene segment for splits between regions combined with time periods, cut-off for graph is 0.95. Top scoring examples are in Table XII.

TABLE IX

Top scoring (above 0.85) region specific changes of NA gene segment. Regions with gaps were included. Even though regions with gaps are not visible, they were present in sequences that had too few representatives, excluding them would leave only positions 328, 504, 512, 1191, 1249, 1300.

Position 94 Score 0.98	Eurasia	Americas	Position 201 Score 0.96	Eurasia	Americas
T C	0.9 0.088	0.995 0.001	— A	0.299 0.675	0.002 0.983
Position 328 Score 0.91	Eurasia	Americas	Position 504 Score 0.96	Eurasia	Americas
T C	0.009 0.983	0.087 0.913	T C	0.857 0.141	0.995 0.004
Position 512 Score 0.92	Eurasia	Americas	Position 1191 Score 0.85	Eurasia	Americas
AG CG	0.991 0.009	0.894 0.106	AG TC AA GA	0.81 0.025 0.095 0.062	0.951 0.034 0.009 0
Position 1249 Score 0.89	Eurasia	Americas	Position 1300 Score 0.96	Eurasia	Americas
A C	0.013 0.984	0.1 0.9	GGTC CGTC	0.903 0.064	0.975 0
Position 1493 Score 0.94	Eurasia	Americas	Position 1497 Score 0.83	Eurasia	Americas
— GG	0.086 0.901	0.007 0.99	— C	0.09 0.907	0.008 0.992

TABLE X

Top scoring (above 0.85) time period specific changes of NA gene segment. If regions with gaps were excluded only positions 340, 507, 698, 548 and 1184 would remain.

Position 201 Score: 0.9	Until 1999	2000–2004	2005–2009	Position 340 Score: 0.92	Until 1999	2000–2004	2005–2009	Position 507 Score: 0.99	Until 1999	2000–2004	2005–2009
—	0.005	0.116	0.258	A	0.848	0.951	0.967	A	0.998	0.999	0.882
A	0.959	0.86	0.733	C	0.117	0.03	0.005	G	0.001	0.001	0.117
Position 548 Score: 0.92	Until 1999	2000–2004	2005–2009	Position 698 Score: 0.88	Until 1999	2000–2004	2005–2009				
AG	0.005	0.104	0.233	AC	0.073	0.026	0.031				
GG	0.974	0.891	0.76	GA	0.925	0.97	0.859				
				GG	0.001	0.001	0.107				
Position 1184 Score: 0.85	Until 1999	2000–2004	2005–2009	Position 1558 Score: 0.88	Until 1999	2000–2004	2005–2009				
AACTAAAAGCATTAGTTCAAGAAAC	0	0	0.06	TTCTACT	0.074	0.052	0.01				
AACCAAAAGCACTAATTCCAGGAGC	0.003	0.107	0.159	TCTACT-	0.039	0.0432	0.008				
AACGATCAGCGAGAAGTTACGCTCA	0.101	0.389	0.097	—	0.845	0.850	0.961				
AACTAAAAGTAACAGACTTAGAAAG	0	0.001	0.068	TCTAC–	0.025	0.04	0.009				
GACTAAAAGTAACAGACTTAGAAAG	0.007	0.082	0.016								
AACGATCAGCGAGAAGTCACGCTTA	0	0	0.1								
GACCAAAAGTAACAGACTTAGAAAG	0	0.001	0.094								

TABLE XI

Most interesting of the top scoring (above 0.85) host specific changes of NA gene segment, all in all 208 such places were found. If regions with gaps are ignored 127 positions would remain, of these presented: 328, 533, 693, 755, 845, 990, 1084 (as a subpart of 1078) and 1384.

Position 96 Score 0.86	Swine	Avian	Equine	Human Score 0.96	Position 151	Swine	Avian	Equine	Human
GGAATGGC-TAACTTAATATTACAA	0.002	0	0	0.045	A	0.145	0.273	0.101	0.007
GGGATAAT-TAGTCTAATGTTGCAA	0.002	0	0	0.04	T	0.009	0.113	0.876	0
TCCACAAT-ATGCTTCTTCATGCAA	0.004	0	0	0.217	G	0.844	0.597	0	0.992
GCAACAGT-ATGTTTCCTCATGCAG	0.002	0.079	0	0					
GGAATAGT-TAGCTTAATGTTACAA	0.004	0.196	0	0.029					
GGAATAAT-TAGTCTAATGTTGCAA	0.002	0.001	0	0.15					
GCCACAAT-ATGCTTCCTTATGCAA	0.067	0.002	0	0.269					
Position 171 Score 0.86	Swine	Avian	Equine	Human Score 0.88	Position 265	Swine	Avian	Equine	Human
—ATTCAACTCCCCCCC-AAA–	0.004	0	0	0.23	—	0.002	0.058	0	0.008
—AGTCAAAACCACACTGGAA–	0.013	0.001	0	0.149	AACAC	0.388	0.02	0	0.943
—AATCAACACCAAGCTGAAC–	0.007	0.105	0	0.006	AATGT	0.002	0.074	0.011	0
—AATCAACGCCAAGCTGAAC–	0	0.062	0	0.003	AGCAC	0.022	0.08	0	0
—AATCAAAATCAGATTGAAA–	0.002	0	0	0.046	AATAC	0.414	0.488	0.427	0.042
—AGTCAAAACAACACTGGAA–	0	0	0	0.068	GAACC	0	0.042	0	0
—ATGCAACTCCCCCCC-AAA–	0.052	0	0	0.24	AACAA	0.03	0.051	0	0
Position 328 Score 0.96	Swine	Avian	Equine	Human Score 0.99	Position 386	Swine	Avian	Equine	Human
T	0.002	0.104	0.011	0	A	0.013	0.162	0.101	0.001
C	0.998	0.884	0.978	1	C	0.987	0.838	0.899	0.999
Position 533 Score 0.98	Swine	Avian	Equine	Human Score 0.94	Position 657	Swine	Avian	Equine	Human
TT	0.056	0.29	0	0.001	A	0.937	0.698	0.798	0.989
AT	0.87	0.682	1	0.991	C	0.022	0.064	0.202	0
					G	0.041	0.236	0	0.012

Position 693 Score 1	Swine	Avian	Equine	Human Score 0.86	Position 815	Swine	Avian	Equine	Human
T	0.991	0.85	0.1	1	A	0.022	0.172	0.809	0
C	0.009	0.15	0.9	0	T	0.94	0.65	0	0.963
					C	0.035	0.17	0.112	0.036
Position 755 Score 0.95	Swine	Avian	Equine	Human Score 0.87	Position 818	Swine	Avian	Equine	Human
TGC	0.247	0.517	0.9	0.006	AAC	0.002	0.051	0	0.001
TGT	0.753	0.482	0.1	0.994	CAC	0.095	0.21	0.1	0.003
					TAC	0.85	0.301	0	0.956
					CAT	0	0.078	0	0.011
					TAT	0.043	0.272	0.9	0.03
Position 845 Score 0.95	Swine	Avian	Equine	Human Score 0.88	Position 903	Swine	Avian	Equine	Human
AG	0.057	0.203	0	0.001	GAAGA	0.082	0.207	0.112	0.017
AA	0.061	0.248	0.1	0.02	GAGGA	0.915	0.792	0.888	0.983
GA	0.872	0.444	0	0.978					
CG	0.009	0.06	0	0					
Position 990 Score 0.93	Swine	Avian	Equine	Human Score 0.85	Position 1008	Swine	Avian	Equine	Human
A	0.033	0.324	0.101	0.017	A	0.881	0.557	0.101	0.964
C	0.061	0.17	0.876	0.001	C	0.013	0.088	0.899	0.001
G	0.907	0.444	0	0.981	G	0.106	0.325	0	0.035
T	0	0.063	0.022	0.001					

Position 1078 Score 0.85	Swine	Avian	Equine	Human Score 0.88	Position 1117	Swine	Avian	Equine	Human
GCTTCAGCA—GTAGCCATTGCCT	0.024	0.002	0	0.108	AG	0.002	0.07	0.112	0
GAGAGGGCA—GCTGTAATCCAGT	0	0.001	0	0.052	GA	0.959	0.362	0	0.96
GCTCCAGCA—GTAGCCATTGTTT	0	0	0	0.072	GG	0.026	0.475	0	0.04
GCTCCAGCA—GTAGCCATTGCCT	0.017	0	0	0.07	GT	0.009	0.073	0	0
GCTCCAGCA—GTAGCCATTGCTT	0.041	0	0	0.234					
GAACAGGTA—GTTGTGGTCCGGT	0.007	0.184	0	0.019					
GAAAGGGCA—GCTGTAATCCAGT	0.002	0	0	0.073					
GAGAGGGCA—GCTGCAATCCAGT	0	0	0	0.0816					
AGACAGGCA—GTTGTGGTCCAGT	0.002	0	0	0.046					

Position 1299 Score 0.93	Swine	Avian	Equine	Human Score 0.97	Position 1319	Swine	Avian	Equine	Human
A	0.408	0.007	0	0.6	TTT	0.883	0.701	1	0.996
T	0.577	0.986	1	0.381	CTT	0.115	0.207	0	0.004
					ATT	0	0.06	0	0

Position 1384 Score 0.94	Swine	Avian	Equine	Human Score 0.97	Position 1433	Swine	Avian	Equine	Human
AT	0.989	0.989	1	0.751	CAG	0.887	0.584	0.101	0.993
GT	0.011	0.006	0	0.248	TAG	0.104	0.341	0	0.004
					CTC	0.004	0.077	0.9	0

TABLE XII

Top scoring (above 0.85) time period and region specific changes of NA gene segment. If regions with gaps are ignored only 319, 328, 403, 408, 457, 497, 504, 507, 512, 548, 555, 560, 626, 677, 698, 701, 704, 761 and 788 would remain.

Position 9 Score: 0.96	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 46 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
————————	0.73	0.788	0.77	0.706	0.622	0.932	—	0.061	0.049	0.024	0.003	0.018	0.052
————————CAAAAGCA	0.009	0.001	0	0.022	0.148	0.004	AATCCAAACCA	0.014	0.042	0.035	0.007	0.013	0.107
————————AGCAAAAGCAG	0.041	0.079	0.085	0.027	0.019	0.011	AACCCAAATCA	0.009	0.025	0.057	0.004	0.006	0.127
————————AGCAAAAGCA	0.092	0.05	0.038	0.084	0.058	0.006	AATCCAAATCA	0.879	0.774	0.788	0.875	0.869	0.644
————————AGCAAAAGC	0.016	0.019	0.004	0.057	0.021	0.017							
Position 62 Score: 0.85	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 72 Score: 0.85	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
-TAATAGC	0.103	0.193	0.087	0.03	0.007	0.002	TCGG	0.006	0.2	0.271	0.049	0.044	0.025
-TAATAAC	0.773	0.677	0.853	0.864	0.9	0.905	TTGG	0.906	0.747	0.694	0.855	0.884	0.898
							TCTC	0.008	0	0.001	0.043	0.053	0.042
Position 77 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 94 Score: 0.99	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
-GC	0.032	0.02	0.019	0.067	0.032	0.047	T	0.977	0.874	0.897	0.995	0.994	0.995
-GT	0.036	0.033	0.007	0.038	0.051	0.018	C	0.001	0.116	0.1	0.002	0	0.001
-TC	0.887	0.927	0.954	0.778	0.825	0.906							
Position 156 Score: 0.85	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 158 Score: 0.87	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.11	0.071	0.02	0.079	0.051	0.024	AA	0.936	0.946	0.963	0.868	0.943	0.852
C	0.007	0.004	0.009	0.053	0.012	0.036	AG	0.009	0.007	0.012	0.014	0.019	0.093
T	0.853	0.899	0.960	0.851	0.918	0.923	GA	0.02	0.035	0.011	0.029	0.01	0.018
Position 171 Score: 0.94	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 178 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
—	0.957	0.98	0.982	0.913	0.936	0.945	AA	0.017	0.003	0.007	0.124	0.067	0.037
CACC	0.022	0.012	0.005	0.041	0.021	0.032	CA	0.620	0.862	0.943	0.605	0.757	0.867
							CC	0.043	0.073	0.011	0.039	0.032	0.007
							CG	0.205	0.014	0.007	0.083	0.002	0.005
							GG	0.018	0.009	0.009	0.047	0.012	0.034
Position 192 Score: 0.89	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 201 Score: 0.99	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AA–	0.612	0.538	0.428	0.626	0.793	0.888	—	0.006	0.217	0.497	0.005	0.002	0
AC–	0.09	0.278	0.38	0.007	0.002	0.006	A	0.952	0.748	0.492	0.973	0.983	0.992
AG–	0.13	0.021	0.011	0.089	0.044	0.001							
AT–	0.025	0.017	0.086	0.001	0	0							
GA–	0.036	0.032	0.01	0.064	0.024	0.009							

Position 202 Score: 0.88	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 262 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
—	0.871	0.933	0.961	0.778	0.857	0.869	—	0.126	0.083	0.018	0.093	0.076	0.044
G	0.016	0.009	0.005	0.039	0.053	0.039	A	0.856	0.894	0.963	0.884	0.876	0.868
T	0.075	0.045	0.018	0.119	0.062	0.0589	T	0	0	0.008	0.023	0.044	0.085
Position 271 Score: 0.96	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 319 Score: 0.87	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
—	0.09	0.067	0.005	0.029	0.023	0	GA	0.042	0.012	0.014	0.098	0.03	0.052
A	0.887	0.924	0.984	0.934	0.932	0.975	TC	0.903	0.960	0.958	0.805	0.884	0.898
Position 328 Score: 0.95	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 340 Score: 0.94	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
C	0.967	0.984	0.989	0.898	0.932	0.916	A	0.912	0.949	0.969	0.818	0.936	0.96
T	0.015	0.011	0.005	0.102	0.068	0.084	C	0.071	0.037	0.009	0.14	0.025	0.001
Position 403 Score: 0.96	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 408 Score: 0.88	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AT	0.004	0.005	0.01	0.04	0.043	0.089	AAG	0.686	0.938	0.744	0.79	0.919	0.961
GT	0.996	0.991	0.99	0.96	0.957	0.911	CAG	0.014	0.006	0.009	0.051	0.065	0.018
							GAG	0.285	0.029	0.015	0.146	0.012	0.013
							TAG	0.012	0.024	0.229	0.005	0	0.001
Position 433 Score: 0.85	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 457 Score: 0.96	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.045	0.041	0.021	0.12	0.152	0.121	TT	0.148	0.363	0.549	0.023	0.002	0.106
C	0.946	0.947	0.979	0.877	0.847	0.879	CT	0.852	0.636	0.450	0.974	0.998	0.894
CT	0.852	0.636	0.450	0.974	0.998	0.894							
TT	0.148	0.363	0.549	0.023	0.002	0.106							
Position 497 Score: 0.98	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 504 Score: 0.99	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.002	0.006	0.01	0.023	0.043	0.09	C	0.083	0.222	0.092	0.006	0.001	0.004
G	0.998	0.994	0.989	0.976	0.956	0.909	T	0.906	0.778	0.907	0.993	0.999	0.995
Position 507 Score: 0.996	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 512 Score: 0.95	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.998	0.999	0.96	0.997	0.998	0.787	AG	0.978	0.994	0.993	0.919	0.975	0.836
G	0	0.001	0.039	0.002	0.002	0.213	CG	0.022	0.006	0.006	0.081	0.024	0.164
Position 548 Score: 0.93	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 555 Score: 0.99	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AG	0.009	0.06	0.155	0.004	0.133	0.285	C	0.955	0.991	0.969	0.999	0.999	0.9
GG	0.982	0.937	0.84	0.972	0.856	0.704	T	0.045	0.009	0.031	0.001	0.001	0.1

Position 560 Score: 0.87	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 626 Score: 0.95	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AA	0.042	0.012	0.014	0.098	0.03	0.052	AGA	0.032	0.003	0.009	0.045	0.02	0.035
AG	0.022	0.012	0.007	0.043	0.021	0.035	TGG	0.955	0.98	0.982	0.913	0.936	0.945
GT	0.044	0.035	0.008	0.041	0.053	0.027							
TC	0.834	0.921	0.953	0.772	0.833	0.863							
Position 677 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 698 Score: 0.91	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.925	0.966	0.968	0.832	0.917	0.893	AC	0.042	0.017	0.013	0.087	0.064	0.055
C	0.029	0.014	0.012	0.071	0.019	0.052	GA	0.955	0.978	0.906	0.912	0.935	0.805
T	0.044	0.02	0.018	0.087	0.064	0.055	GG	0.001	0.002	0.077	0.001	0	0.14
Position 701 Score: 0.87	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 704 Score: 0.92	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.898	0.939	0.972	0.826	0.898	0.878	AT	0.98	0.977	0.988	0.971	0.983	0.898
G	0.058	0.026	0.02	0.135	0.051	0.099	GT	0.017	0.021	0.01	0.021	0.013	0.096
T	0.044	0.035	0.008	0.039	0.051	0.0237							
Position 761 Score: 0.99	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 788 Score: 0.95	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AA	0.929	0.951	0.977	0.897	0.907	0.932	GATGG	0.923	0.822	0.877	0.906	0.994	0.993
CA	0.021	0.01	0.009	0.047	0.011	0.035	GACGG	0.076	0.176	0.122	0.093	0.006	0.007
GA	0.038	0.036	0.011	0.013	0.04	0.0292							
Position 819 Score: 0.85	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 830 Score: 0.89	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AC	0.893	0.698	0.482	0.914	0.929	0.966	AT	0.983	0.976	0.977	0.881	0.899	0.922
AT	0.063	0.267	0.508	0.047	0.019	0.01	GT	0.015	0.018	0.02	0.069	0.09	0.043
CA	0.044	0.035	0.008	0.039	0.051	0.022							
Position 854 Score: 0.9	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 865 Score: 0.89	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
-AA	0.891	0.979	0.952	0.857	0.888	0.95	AG	0.021	0.035	0.007	0.039	0.049	0.023
-AG	0.094	0.009	0.021	0.105	0.071	0.025	A-	0.913	0.948	0.976	0.88	0.924	0.941
							G-	0.043	0.015	0.016	0.081	0.025	0.035
Position 876 Score: 0.85	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 882 Score: 0.85	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
—	0	0.103	0.0968	0	0	0	A	0.896	0.754	0.574	0.759	0.939	0.956
A	0.944	0.845	0.857	0.907	0.898	0.949	G	0.09	0.239	0.418	0.2	0.043	0.014
G	0.048	0.026	0.011	0.053	0.059	0.03							
T	0.009	0.024	0.035	0.025	0.011	0.012							

Position 942 Score: 0.88	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.908	0.952	0.974	0.793	0.887	0.914
G	0.048	0.013	0.018	0.168	0.062	0.062
T	0.044	0.035	0.008	0.039	0.051	0.024

Position 975 Score: 0.88	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AC	0.039	0.009	0.011	0.081	0.023	0.036
GC	0.034	0.017	0.014	0.063	0.053	0.044
TC	0.921	0.966	0.972	0.831	0.913	0.898

Position 1002 Score: 0.85	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AA	0.893	0.957	0.963	0.747	0.886	0.923
GA	0.056	0.026	0.018	0.164	0.09	0.041
TC	0.039	0.009	0.011	0.081	0.023	0.04

Position 1053 Score: 0.85	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AC	0.065	0.023	0.022	0.137	0.082	0.072
AG	0.022	0.01	0.009	0.049	0.012	0.035
GG	0.869	0.933	0.961	0.776	0.855	0.869
TC	0.044	0.0347	0.008	0.039	0.051	0.023

Position 1084 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.061	0.046	0.011	0.049	0.042	0.024
G	0.882	0.938	0.97	0.81	0.886	0.903
T	0.056	0.016	0.018	0.141	0.062	0.072

Position 1086 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A—	0.912	0.966	0.967	0.806	0.9	0.886
G—	0.044	0.021	0.018	0.088	0.068	0.056

Position 1137 Score: 0.99	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AT	0.009	0.049	0.088	0.006	0.001	0.005
GT	0.991	0.951	0.911	0.993	0.999	0.995

Position 1161 Score: 0.92	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GA	0.015	0.015	0.051	0.007	0.004	0.084
GG	0.975	0.983	0.947	0.977	0.963	0.907

Position 1164 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AATG	0.731	0.95	0.941	0.709	0.818	0.816
GATG	0.167	0.01	0.017	0.131	0.076	0.104

Position 1173 Score: 0.87	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
TGGAT	0.89	0.94	0.969	0.824	0.863	0.904
TGGCT	0.048	0.02	0.019	0.042	0.02	0.012
TGGTT	0.041	0.035	0.007	0.084	0.098	0.067

Position 1182 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AGAACTAAAAGTAACAGACTTAGAA	0	0	0.001	0	0	0.149
AGAACCAAAAGCACTAATTCCAGGA	0.01	0.201	0.303	0	0	0
AGGACTAAAAGTAACAGACTTAGAA	0.012	0.0307	0.02	0	0.09	0.004
AGGACCAAAAGTAACAGACTTAGAA	0	0.001	0.07	0	0	0.118
AGAACGATCAGCGAGAAGTCACGCT	0.021	0.005	0.032	0.016	0.004	0.215
AGAACAATCAGTAGGGATTCGCGGT	0	0	0	0.004	0.036	0.082
AGAACTAAAAGCATTAGTTCAAGAA	0	0	0.024	0	0	0.1
AGAACGATCAGCGAGAAGTTACGCT	0.12	0.223	0.105	0.087	0.464	0.064

Position 1196 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GGATTCGCGGTCAGGCTATGAGACA	0	0	0	0.004	0.036	0.079
GAAGTTACGCTCAGGATATGAAACC	0.061	0.15	0.1	0.02	0.201	0.046
GGATTCACGCTCAGGTTATGAGACT	0.041	0.07	0.0186	0	0	0
GAAGTTACGCTCAGGTTATGAAACC	0.059	0.082	0.005	0.082	0.261	0.009
TAATTCCAGGAGCGGCTTTGAAATG	0.011	0.25	0.388	0	0	0
GGATTCGCGCTCAGGTTATGAGACT	0.015	0.104	0.035	0	0	0
TAGTTCAAGAAACGGTTTTGAGATG	0	0.027	0	0	0	0.1
GAAGTCACGCTTAGGGTATGAAACC	0	0.015	0	0	0	0.202
CAGACTTAGAAAGGGGTTTGAGATG	0.0123	0.031	0.078	0	0.084	0.27

Position 1239 Score: 0.93	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GAA	0.022	0.012	0.006	0.038	0.014	0.035
TGG	0.955	0.98	0.981	0.912	0.936	0.945

Position 1249 Score: 0.94	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.032	0.007	0.011	0.14	0.055	0.086
C	0.96	0.992	0.988	0.86	0.945	0.914

Position 1252 Score: 0.88	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 1300 Score: 0.98	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.909	0.95	0.975	0.839	0.909	0.913	CGTC	0.016	0.117	0.036	0	0	0
C	0.036	0.035	0.007	0.037	0.051	0.023	GGTC	0.923	0.845	0.946	0.959	0.988	0.982
G	0.055	0.015	0.018	0.124	0.04	0.064							
Position 1317 Score: 0.87	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 1331 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
A	0.915	0.968	0.972	0.834	0.938	0.899	A	0.021	0.009	0.011	0.031	0.015	0.039
G	0.032	0.017	0.009	0.06	0.029	0.053	C	0.045	0.047	0.024	0.031	0.033	0.006
T	0.05	0.015	0.019	0.105	0.033	0.047	G	0.021	0.002	0.002	0.075	0.017	0.019
							T	0.913	0.942	0.963	0.864	0.935	0.937
Position 1460 Score: 0.88	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 1464 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
AG	0.865	0.918	0.952	0.814	0.882	0.86	A	0.896	0.925	0.966	0.838	0.885	0.915
CA	0.065	0.039	0.012	0.119	0.074	0.06	C	0.034	0.021	0.012	0.058	0.021	0.035
GG	0.029	0.01	0.006	0.047	0.029	0.05	G	0.052	0.036	0.016	0.075	0.051	0.029
Position 1469 Score: 0.9	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 1490 Score: 0.99	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
GG	0.017	0.015	0.01	0.037	0.021	0.036	—	0.02	0.134	0.061	0.001	0.001	0.012
TG	0.957	0.958	0.973	0.91	0.93	0.933	GA	0.979	0.866	0.938	0.997	0.998	0.988
Position 1493 Score: 0.97	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09	Position 1497 Score: 0.98	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09
—	0.027	0.142	0.062	0.003	0.001	0.013	—	0.027	0.143	0.068	0.004	0.002	0.013
GG	0.951	0.854	0.921	0.995	0.990	0.987	C	0.961	0.855	0.932	0.994	0.998	0.987
Position 1546 Score: 0.86	EuA u 99	EuA 00–04	EuA 05–09	As u 99	As 00–04	As 05–09							
AAAACTCCTTGTTTCTACT——	0.036	0.025	0.004	0.058	0.115	0.007							
———————	0.636	0.743	0.841	0.668	0.555	0.882							
AAACTCCTTGTTTCTAC———	0.074	0.074	0.018	0	0.001188	0.001							
AAACTCCTTGTTTCTACT——	0.047	0.016	0.01	0.035	0.129	0.006							
GTTTCTACT—————	0.031	0.039	0.042	0.012	0.005	0.002							
TGTTTCTACT—————	0.018	0.021	0.016	0.042	0.0167	0.016							

4.5. Column entropy

We also looked at how variable all of our gene segments were. To do this we calculated the entropy of each site in the alignment. Sites with gaps were excluded. The formula used to estimate the entropy for the j-th column is naturally,

$$\hat{I}_j = -\sum_{R\in\{A,C,G,T\}} \frac{n_R}{n} \log \frac{n_R}{n}\,, \tag{1}$$

where n is the amount of sequences, and n_R is the count of the given residue in the considered column. This famous formula can be thought to represent how "random" the column is. If the column is constant it will be 0 and if all residues are equally present (*i.e.* for all R $n_R/n = 0.25$) then it will obtain its maximum of $\log 4$. The results are in Figs. 6, 7 and 8.

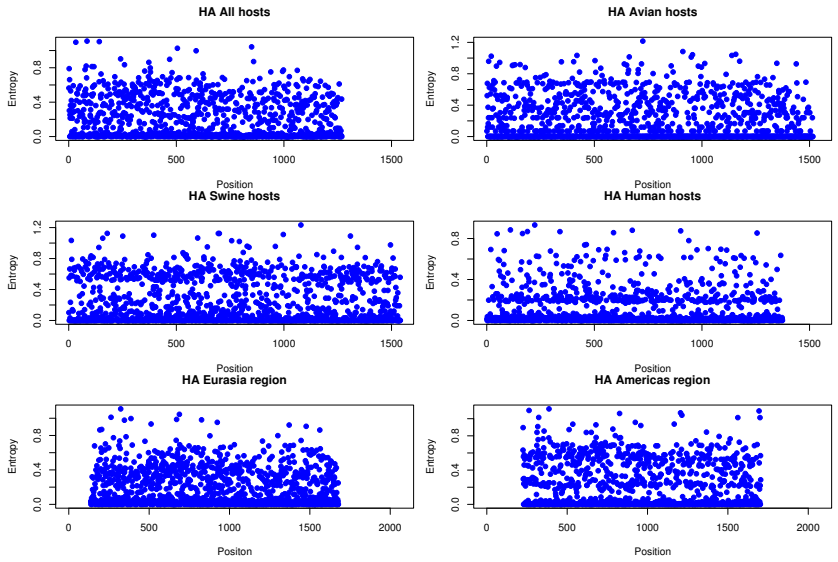

Fig. 6. Estimated entropies at individual sites in alignment of HA gene segment sequences and compared with breaking the sequences according to hosts and geographical regions.

What can be seen from the figure is that there is not that much difference between hosts and regions in each gene segment but there are differences between the gene segments. This could be because of different selective pressures on the different gene segments due to the M protein being an internal protein while the HA and NA are external proteins. External proteins in viruses are more liable for change as they are responsible for the virus having the ability to infect. Therefore the difference between gene segments is rather obvious but the lack of difference between hosts and regions could be interesting for further biological and modelling insight.

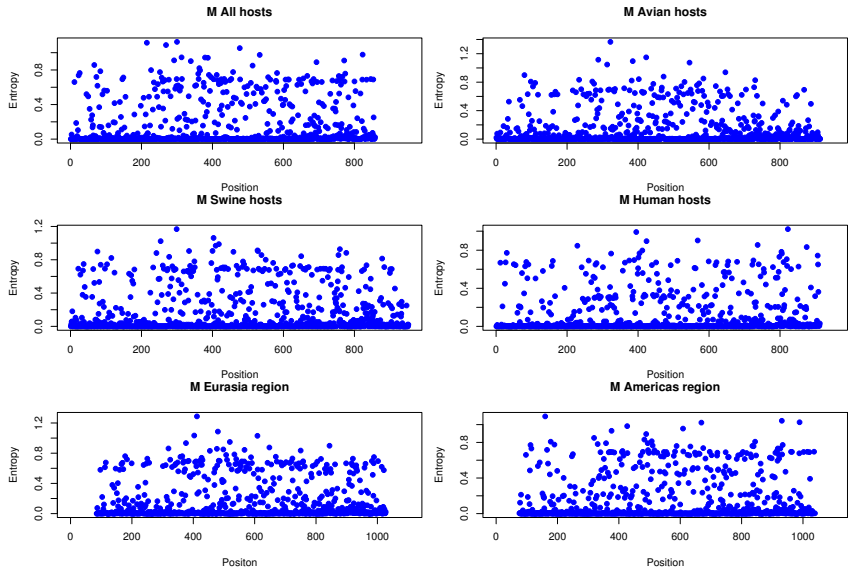

Fig. 7. Estimated entropies at individual sites in alignment of M gene segment sequences and compared with breaking the sequences according to hosts and geographical regions. Entropy not calculated for columns with gaps.

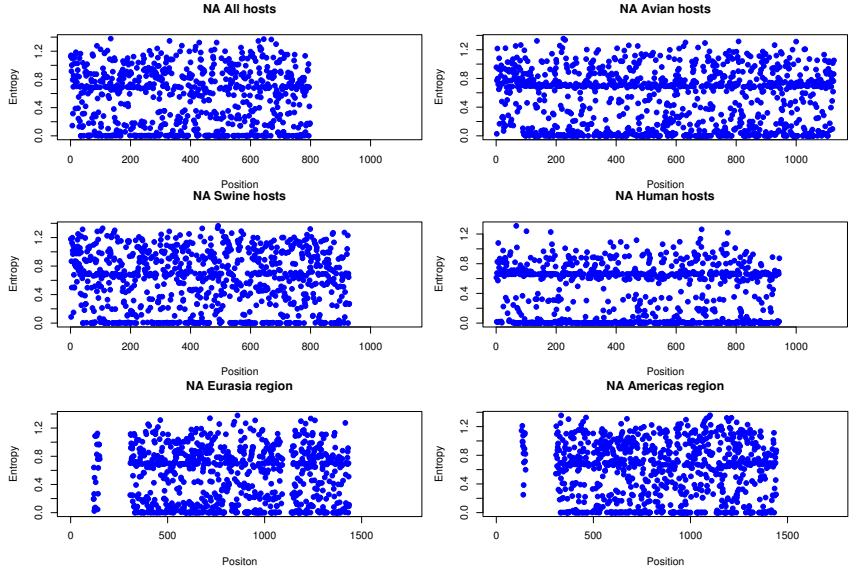

Fig. 8. Estimated entropies at individual sites in alignment of NA gene segment sequences and compared with breaking the sequences according to hosts and geographical regions. Entropy not calculated for columns with gaps.

5. Discussion

Our aim was to using a simple methodology to find regions of up to 25 base pairs in influenza viral gene segment sequences that differentiate between certain factors (host, geography, time period, Hx strain). We wanted to see whether anything novel could be found using a rather "rough look" at data. The aim was achieved in a sense that a number of promising places in the gene segments were found. Because the methodology is very simple and requires some heuristic scoring mechanism its drawbacks have to be discussed.

5.1. Sequence acquisition

The sequences were downloaded from GISAID [4]. Therefore the methodology is dependent on us obtaining a random sample of sequences which when looking at the sequence breakdown can be seen to be immediately violated. It is clear that the majority of sequences have been collected from the past ten years and from human hosts. Another issue is that sometimes the lengths of the downloaded sequence were drastically different. For example in the HA gene segment of H1N1 about 1000 sequences were less than half the length of the others (so one third of the sequences). They had to be discarded as they made the alignment unsure and also the program was finding only differences between these two groups *i.e.* gaps/no gaps. It cannot be assessed therefore whether this could have caused any bias. One issue with the downloaded sequences is that we do not have all the information for every single one, sometimes region, host or time is missing. This then cuts down the sample sizes in some parts of the study.

5.2. Method

The first step of the analysis is to align the acquired sequences. This was done in `Clustal 10.2` [5]. Due to the huge number of sequences (a couple of thousand in each case) the alignment has to be done approximately. Visual inspection of it showed extremely high similarity between all of the sequences and no clear misaligned regions could be seen. But this is the grand picture, small regions/individual bases could be very well misaligned which can have profound effects on the method of searching for characteristic differences. It was noticed very often that in most places one had a huge number (around 200) of different values of the metasubsequence that had one, two, less than ten sequences. This could be due to mutations in those individual sequences (which could very well be important) but just as well due to a misalignment. Another observation made is that especially in the M and NA gene segments nearly every single metasubsequence of length 25 was seen by the

contingency table test (at a p-value of 1^{-183}) to be significant in the spatial and temporal analysis. Such a huge number of positions has to be cut down naturally. At the moment this is done first by only considering those values of the metasubsequence that have at least 2% of the sequences altogether and then assigning scores to each contingency table and we are left with the presented results. Because of the non-independence the p-values should be treated as a cut-off value and not understood in the statistical sense. A last issue is how to combine factors, *e.g.* countries, time — periods. The division of time-periods should be sequential. Here we divided the time into three periods to keep the number of sequences in each bin approximately the same. But this might cause us to miss some effects inside these bins and there also comes the potential problem should the bins be the same for every geographical region. An issue in the creation of categories is computational, for this to be feasible there cannot be too many, we considered a maximum of 12 categories (with approximately 10000 sequences). Our analysis does not address these issues. It is not its aim. The aim is to find as many as possible potentially interesting sites in the flu gene segments that show differentiation between hosts, geography and time by a simple first glance.

5.2.1. Scoring method

To cut down on the huge number of significant tables we devised a scoring method described in Section 3.3. The intuition behind it is that it will highly score those tables where given the group we get a lot of information about the sequence. An obvious issue with the method is to decide the cut-off value. At the moment we choose a level such that the number of results would be manageable to look through. When looking at Table XIII we can see that positions 201 and 202 are in it separately. This is because together they did not generate a significant contingency table. Something else that can be noticed in the presented results is that they show positions which start and end with constant columns. These of course do not influence the score but they might play a role in the p-value. We have to remember that to calculate the score we remove those rows that have less than 2% of the sequences. There could be changes on these "constant" positions in the removed rows which decide that the table is significant.

5.3. Evolution model

In a recently published paper [16] the authors present a mathematical model which they claim well describes the creation of viral phylogenetic trees. From our perspective when analyzing the influenza differentiating metasubsequences, the important conclusion from [16] is that one small change can create a virus version which will be present in significant num-

bers. This is strikingly visible (unless this is due to a sampling bias over which we have no control) when looking at the values of the metasubsequence at position 1023 of the M gene segment. We have a version of the virus that in the Americas was very scarce and then recently started to become visible in a huge proportion. When we looked at the contingency tables in every single sequence of positions in the alignment we observed that there could be around 200 realizations in those positions that had 1, 2, 3 below 10 sequences having that value of the metasubsequence. We said that this could be due to a misalignment but in view of the model and the result in the M gene segment they could just as well be real changes that might or might not become dominant. At the moment this was noticed only in one place.

6. Further developments

The analysis done here is a very superficial first glance at differentiation between influenza viruses due to factors like host, geography or time period. Interactions between time and region were also looked at. The next steps should be introducing models of strain evolution, especially to consider the phylogenetic dependence between them.

K.B. is partially funded by the Kungliga Vetenskapsakademien.

REFERENCES

[1] N. Voirin, B. Barret, M.-H. Metzger, P. Vanhems, *Journal of Hospital Infection* **71**, 1 (2009).

[2] T.P. Weber, N.I. Stilianakis, *Journal of Infection* **57**, 361 (2008).

[3] N. Ferguson, *Nature* **446**, 733 (2007).

[4] GISAID Platform, `http://platform.gisaid.org/`

[5] M. Larkin *et al.*, *Bioinformatics* **23**, 2947 (2007).

[6] W. Ewens, G. Grant, *Statistical Methods in Bioinformatics*, Springer, New York 2005.

[7] A. Tamuri, M. dos Reis, A. Hay, R. Goldstein, *Identifying Changes in Selective Constraints: Host Shifts in Influenza*, *PLoS Computational Biology* **5**(11) (2009).

[8] Y. Furuse, A. Suzuki, T. Kamigaki, H. Oshitani, *Evolution of the M Gene of the Influenza A Virus in Different Host Species: Large Scale Sequence Analysis*, *Virology Journal* **6** (2009).

[9] O. Miotto, A. Heiny, T. Tan, J. August, V. Brusic, *Identification of Human-to-human Transmissibility Factors in PB2 Proteins of Influenza a by Large-scale Mutual Information Analysis*, *BMC Bioinformatics* **9**(Suppl 1) (2008).

[10] W. Pirovano, K. Feenstra, J. Herings, *Nucleic Acids Research* **34**, 6540 (2006).

[11] K. Feenstra, W. Pirovano, K. Krab, J. Herings, *Nucleic Acids Research* **35**, W495 (2007).

[12] Perl v5.8.5., `http://www.perl.com`

[13] J. Stajich *et al.*, *Genome Research* **12**, 1161 (2002).

[14] M. Kospach, Statistics-Distributions-1.02, `http://search.cpan.org/ mikek/Statistics-Distributions-1.02/`

[15] D. Huson, D. Richter, C. Rausch, T.Dezulian, M. Franz, R Rupp, *BMC Bioinformatics* **8**, 460 (2007).

[16] T. Liggett, R. Schinazi, *Journal of Applied Probability* **46**, 601 (2009).