Computational Aspects of In-silico Experiments for Investigating the Impact of the Host Genome on the Influenza Virus A Variability

P. Borovska, V. Gancheva, E. Asenov, I. Georgiev

Key Words: *Bioinformatics; high performance computing; influenza virus; multiple sequence alignment; sequence alignment.*

Abstract. *Nowadays the study of the variability of influenza virus is a problem of very great importance. Influenza type A viruses cause epidemics and pandemics. The problem of restricting the spreading of pandemics and the treatment of the people infected by the influenza virus is widely based on the latest achievements of molecular biology, bioinformatics and biocomputing, as well as many other advanced areas of science. In silico biological sequence processing is a key for molecular biology. This scientific area requires powerful computing resources for exploring large sets of biological data. The paper presents parallel computational simulations for the case study of investigating the role of the host genome in the evolution and fast changeability of the influenza virus A on supercomputer BlueGene/P. The experimental framework is based on all available existing influenza virus A nucleotide sequences, the clustalw algorithm for multiple sequence alignment, the blast algorithm for sequence searching, the Philip software for philogenetic tree reconstruction and the recombination analysis tool for finding hot-spots of mutation/recombination in influenza A virus genomes.*

Introduction

The flu virus A genome consists of eight RNA molecules that replicate in the host cell nucleus. Its replication is realized by RNA-dependent RNA polymerase, which (unlike the DNA polymerase) is devoid of proof reading activity. Due to this it makes mistakes with a frequency of one base substitution per 10 000 nucleotides. Thus any round of replication results in appearing of at least one point mutation per viral genome. Although this frequency is very high (compared to the mutability of the DNA viruses), it cannot satisfactorily explain the extremely fast changeability and adaptively of the flu virus. The latter helps it to escape the immune system, to increase its virulence and to cause unexpected epidemics and pandemics.

Viruses are intracellular molecular parasites and their evolution is closely dependent on the host peculiarities. We assume that both virus and host bear common genetic elements that allow a homologous genetic recombination to occur. As a result specific nucleotide sequences will be exchanged between the host and viral RNAs, which will highly accelerate the process of accumulation of mutations in the viral genome and therefore will speed up its changeability and evolution. Recombination between viral and host genes is observed with many other (both DNA and RNA) viruses, however, it has not been communicated for the flu viruses so far. This hypothesis can be checked experimentally and its proof would explain the (well known) rapid flu virus evolution and adaptivity to new hosts. Identification of mutation and recombination hot-spots in influenza A viral genome will lay the foundations of new approaches for molecular diagnostics and prognostics of flue viral infections as well as for development of new flu strain specific vaccines.

The new technology for full genome sequencing employed after the year 2000 led to accumulation in the GenBank of thousands of influenza viral genome sequences originating from different viral isolates. As the number of DNA and protein sequences databases is increasing, it becomes important to be able to use parallel algorithms for sequence alignments of very large number of sequences. Besides the existing software packages, in most cases the latter are not applicable for the analysis of this information by single or small number processor computers. These problems might be solved by means of utilizing of modern methods of parallel computing employing supercomputers such as the BlueGene.

The Basic Local Alignment Search Tool (BLAST) has been suggested [1,2] and utilizes a heuristics approach for increasing the performance of the alignment searching. BLAST is the most widely used sequence alignments program. BLAST searches a database for sequences similar to other sequence and efficiently calculates local pairwise alignments between sequences. In recent years several parallel BLAST algorithms for alignment search have been reported. [3] introduces the database fragmentation strategy in mpiBLAST. ClustalW [4] implements a progressive method for multiple sequence alignment and is a widely used tool for DNA or proteins. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be found. The algorithm ClustalW proceeds in three steps: pairwise alignment, guide tree and multiple alignment.

The goal of this research is to provide adequate parallel computer simulation platform for investigating the impact of the host genome on the influenza virus variability, including the identification of hot-spots of mutation/recombination in influenza A viral genomes, the investigating of the influenza virus changeability and evolution for predic-

Table 1. Influenza Virus A nucleotide sequences

Segment number	Length (nucleotide bases)	All hosts Influenza Virus A/H1N1		All hosts Influenza Virus A/all subtypes		Human Influenza Virus A/all subtypes		Human Influenza Virus A/H1N1		Avian Influenza Virus A/All Subtypes	
		Isolate count	Input data size (MB)	Isolate count	Input data size (MB)	Isolate count	Input data size (MB)	Isolate count	Input data size (MB)	Isolate count	Input data size (MB)
#1	2310	3780	9.00	10148	24.25	5850	13.94	3430	8.16	5830	12.54
#2	2340	3802	9.07	10607	25.35	5878	14.02	3431	8.18	5876	12.87
#3	2230	3829	8.67	10876	24.73	5904	13.39	3453	7.81	5822	12.16
#4	1780	5992	10.78	19388	29.41	8999	16.15	5441	9.78	9068	14.45
#5	1560	4046	6.45	10894	17.44	6170	9.84	3611	5.75	5539	8.20
#6	1410	6128	9.11	16662	24.73	9662	14.39	5521	8.20	7034	9.89
#7	1030	4567	4.82	12369	13.13	7213	7.65	4109	4.33	6308	6.47
#8	890	3996	3.66	12185	11.20	6150	5.63	3584	3.28	5971	5.37
Segment number	Length (nucleotide bases)	Avian Influenza Virus A/H1N1		Swine Influenza Virus A/All Subtypes		Swine Influenza Virus A/H1N1		Horse Influenza Virus A/H7N7		Horse Influenza Virus A/H3N8	
		Isolate count	Input data size (MB)	Isolate count	Input data size (MB)	Isolate count	Input data size (MB)	Isolate count	Input data size (MB)	Isolate count	Input data size (MB)
#1	2310	131	0.30	654	1.34	332	0.70	8	0.02	95	0.23
#2	2340	133	0.30	684	1.42	329	0.72	7	0.02	97	0.30
#3	2230	134	0.28	651	1.26	321	0.67	7	0.02	89	0.20
#4	1780	152	0.26	1224	1.88	617	0.94	18	0.03	139	0.25
#5	1560	135	0.20	737	1.07	378	0.57	11	0.02	96	0.15
#6	1410	189	0.28	937	1.27	492	0.67	9	0.01	113	0.17
#7	1030	149	0.15	825	0.83	426	0.43	10	0.01	121	0.13
#8	890	127	0.11	732	0.65	365	0.33	15	0.01	113	0.10

tion of influenza epidemics and pandemics, the simulation of the influenza virus interaction with host genome.

2. Computational Framework

2.1. Experimental Platform

The experimental framework includes IBM Blue Gene/P supercomputer, consisting of two racks, 2048 PowerPC 450 based compute nodes, 8192 processor cores and a total of 4 TB random access memory. Double-precision, dual pipe floating-point core acceleration is available on each core. Sixteen I/O nodes are connected via fibre optics to a 10 Gb/s Ethernet switch. The smallest partition size, available currently, is 128 compute nodes (512 processor cores). The maximum LINPACK performance achieved is Rmax = 23.42 Tflops. The theoretical peak performance is Rpeak=27.85 Tflops. Furthermore cupboards with computing nodes supercomputing system include the following major components: 1. Front-End Node: server to which users have access and which put out its tasks. The architecture is PowerPC 64 and Operation System - SuSE Linux Enterprise Server 10 (SLES 10); 2. Service Node (SN): server that manages the overall operation of the system; 3. Two file server by which FEN and computing nodes have to access the shared disk array with 12TB. The Blue Gene/P architecture supports a distributed memory, message-passing programming model. Message passing is based on the MPICH2 distribution of the MPI standard.

2.2. Experimental Database Development

All existing sequences of influenza virus obtained from various isolates are selected. The NCBI Influenza Virus Sequence Database contains nucleotide sequences, protein sequences and their encoding regions of all influenza viruses in GenBank [5], including the complete genome sequences: www.ncbi.nlm.nih.gov. A local database in working format is designed and implemented on the supercomputer BlueGene/P. The local database is a mirror of the existing database and permits online updating of data. This always allows to keep current available database. The local database comprises real datasets of all the available isolates of the 8 segments of the influenza virus A for various hosts, given in *table 1*.

3. Experimental Results and Analisis

3.1. Similarity Searching of Influenza Virus Sequences and Whole Human Genome

The complete human genome is used as a database. The human genome contains 3.4 billion DNA base pairs. The database is segmented into approximately equal sized 64 segments and stored in the shared memory. Different

Table 2. Executing time in the case of avian virus A/H1N1 searching into human genome

Number of queries	852	1196	3331	5951	6729
Execution Time (min)	27,51	42,39	102,22	180,55	201,33

Table 3. Executing time in the case of all hosts influenza virus A/H1N1 nucleotide sequences

Segment number	Isolate count	Length (nucleotide bases)	Input data size (MB)	Execution Time (min)
				2048 cores
#1	3780	2310	9.16	223.59
#2	3802	2340	9.12	283.40
#3	3829	2230	8.76	230.56
#4	5992	1780	10.87	480.27
#5	4046	1560	6.45	302.66
#6	6128	1410	9.11	425.20
#7	4567	1030	4.81	407.58
#8	3996	890	3.65	199.45

data sets of influenza virus nucleotide sequences based on specified criteria such as subtype, segment, host, region, have been used as queries in order to search for similarities with the human genome. All existing sequences of a particular subtype, segment, and host of the influenza virus are combined into a virtual query. This allows comparing of a large set of sequences against a sequence database simultaneously by sending virtual query and reducing the execution time.

In order to satisfy the research purpose the objective of the experiments is the similarity searching of RNA segments of various influenza viruses A/H1N1 strains and the human genome based on sequence alignment method mpiBLAST for the case study of investigating the interaction between the influenza virus A/H1N1 and the host genome.

A number of experiments have been carried out on a supercomputer BlueGene/P utilizing various numbers of virtual queries and data sets. The execution time in the case of avian influenza virus A/H1N1 nucleotide segment 4 (HA) used for searching in the human genome in respect to various virtual queries size and 512 cores are given in the *table 2*. The results show that the batch size impacts the execution time, because more sequences need to be searched.

3.2. Multiple Alignment of Influenza Virus Nucleotide Sequences

Comparisons between RNA segments of various influenza viruses A strains have been carried out based on parallel program MPI-based implementation of ClustalW algorithm for multiple sequence alignment on the supercomputer BlueGene/P. The case study is to investigate the consensus motifs and variable domains. Experiments have been conducted by a parallel MPI-based program implementations on a mirror local database installed on the supercomputer comprising all the available isolates of the 8 segments of the influenza virus A (all subtypes) extracted from NCBI.

A number of experiments comprises nucleotide sequences homology discovery within all the available isolates of the 8 segments of different viruse A (all hosts, human, swine, horse) have been performed. *Table 3* shows some experimental results in the case of all hosts virus A – the segments used for analysis and execution time on 2048 processors. The results show that the sequence length impacts to the execution time.

The molecular biology outcome of the experiments is that the consensus motifs and variable domains in Influenza virus A have been determined and output by utilizing the biological sequence alignment editor UGENE UniPro [6] (*figure 1*).

3.3. Investigation and Visualization of the "Hot" Spots of Recombination of Influenza Virus

Recombination could be the predominant factor in shaping the genome evolution. The Recombination Analysis Tool [7] is intended for high-throughput, distance-based analysis of both DNA and protein multiple sequence alignments. RAT input files are nucleotide or protein sequences and the output is a graphical representation of the points of recombination. The recombination in the case of all horse influenza viruses H3N8, segment 1 is shown in *figure 2*.

The input parameters are: Similarity – 82%; Jumps to over – 92%. The recombination sites of all existing nucleotide and protein sequences of influenza virus, separated by subtype, host and segment after multiple alignment using parallel implementation of ClustalW algorithm have been investigated. The results of recombination analysis in the

Figure 1. Finding out consensus domains in the case of Human Influenza Virus A/H1N1

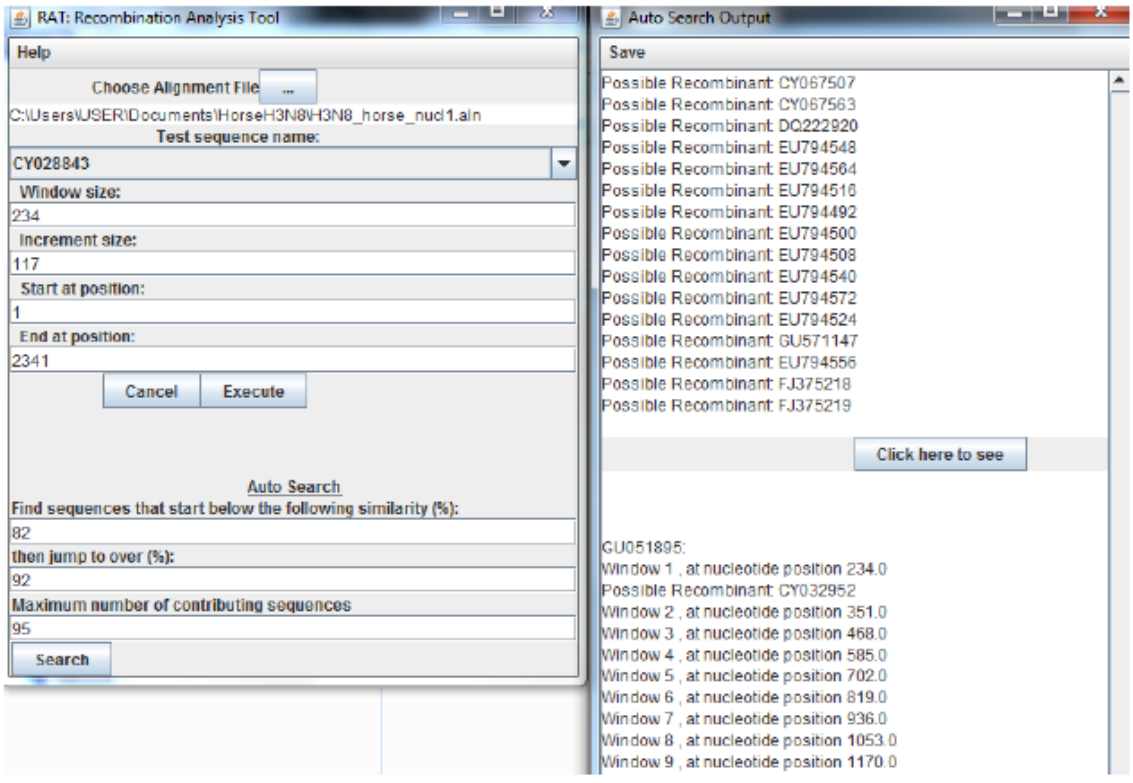

Figure 2. Recombination analysis in the case of all horse influenza viruses H3N8, segment 1

cases of all hosts influenza virus A/H1N1 segment 2 are shown in *figure 3*. Recombination sites have been identified by the RAT to identification of recombination of hot-spots in the influenza virus genome. RAT allows the user to see only those areas of aligned sequence, which is interested.

3.4. Representation of Influenza Virus Phylogenetic Tree

Phylogenetic trees of input sequences are constructed using computational phylogenetic methods [8]. The phylo-

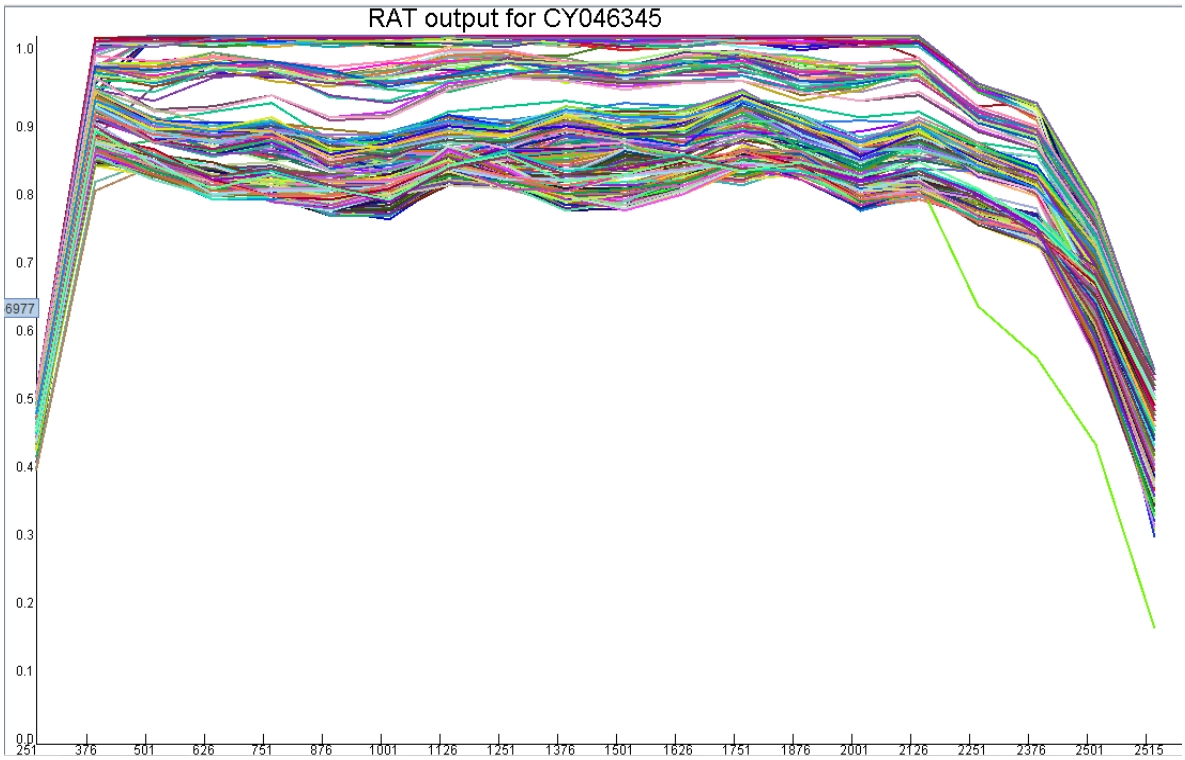

Figure 3. Recombination hot-spots sites in the case of all hosts influenza virus A/H1N1 segment 2

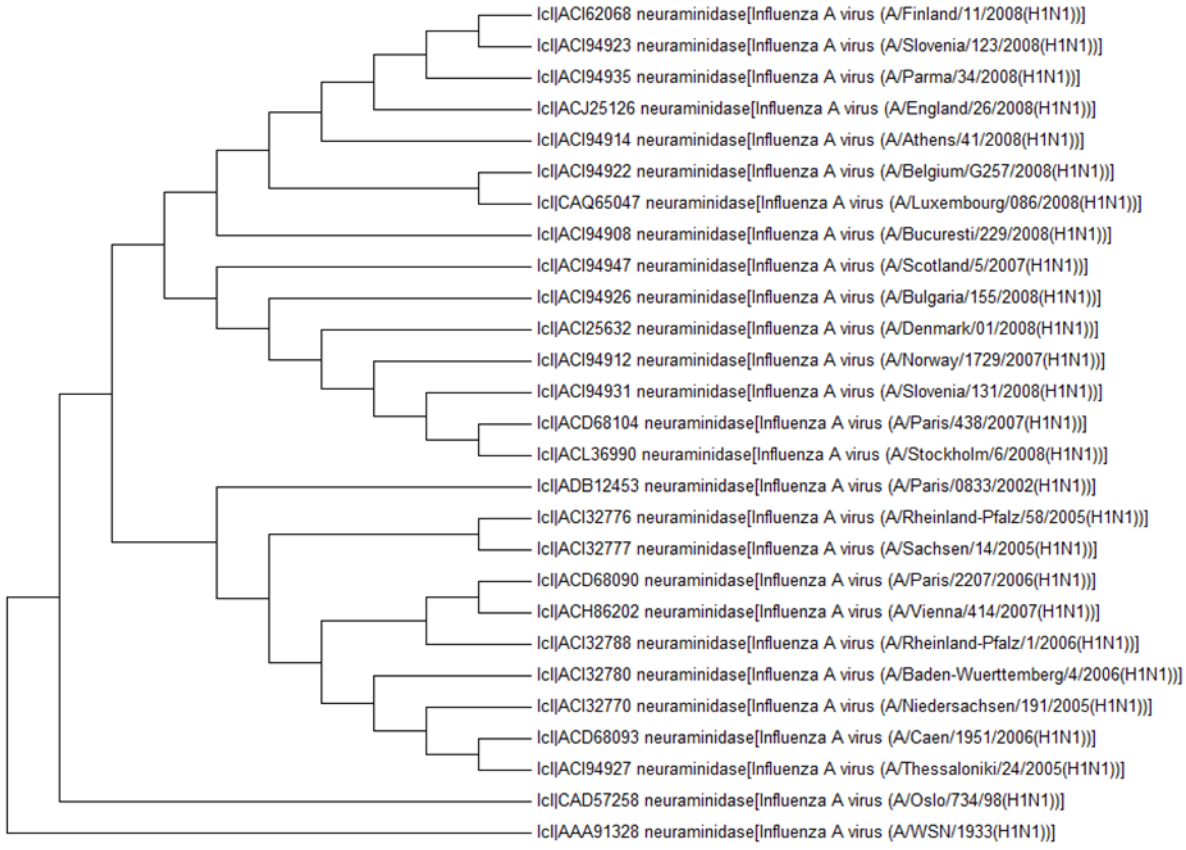

Figure 4. Representation of rectangle phylogenetic trees of NA protein

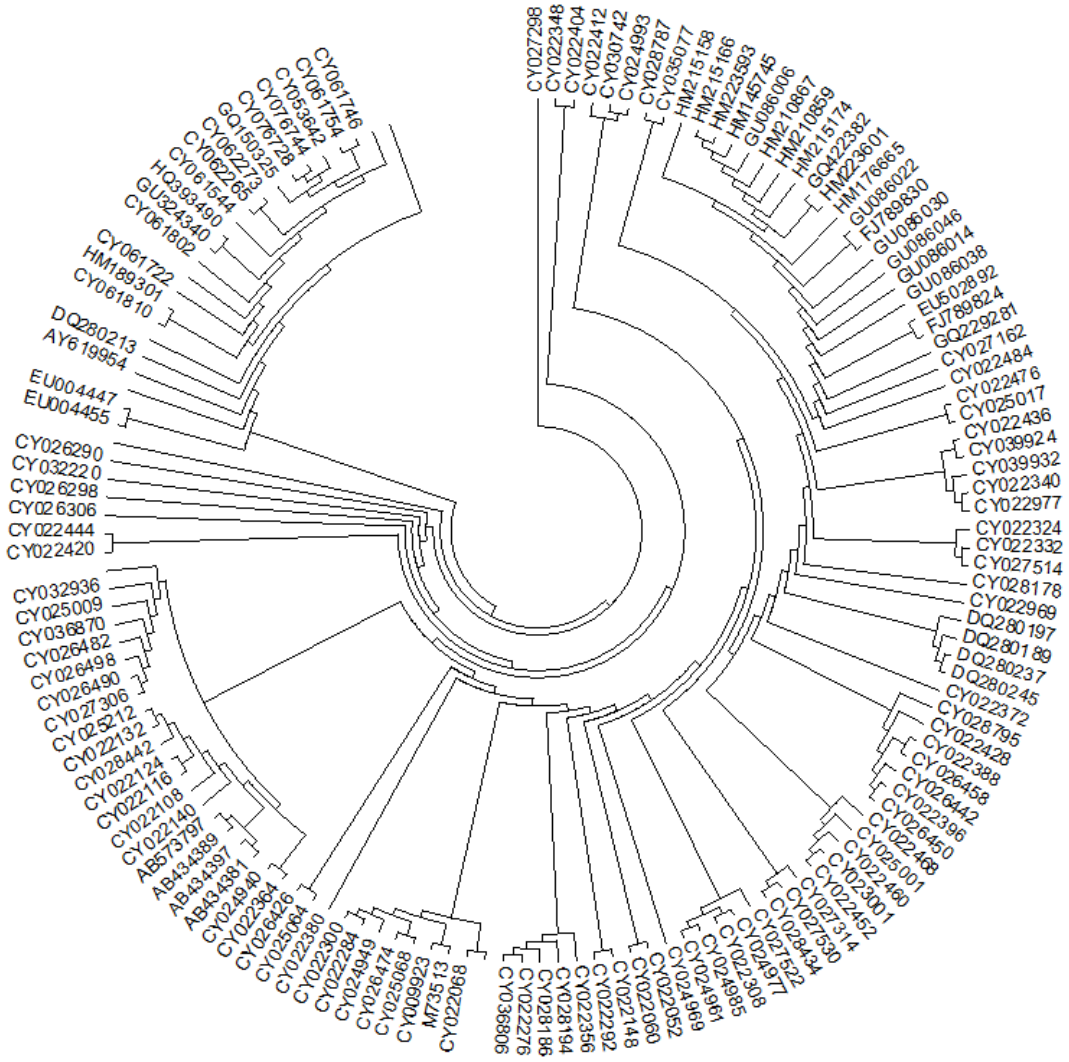

Figure 5. Circular phylogenetic tree of all existing NA protein of influenza virus isolates in human in Europe

genetic tree helps the researchers to show and analyse the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics. The traditional (rectangle) structure of phylogenetic tree in the case of some of protein neuraminidase of an influenza virus, isolated in human is shown in *figure 4*. Using this phylogenetic tree one can establish the relationships between proteins of the influenza virus. For example, the virus that is isolated with an identification code AAA91328 is farthest from virus ACI62068 that is isolated in Finland and the virus ACI94923 that is isolated in Slovenia. So we know that the virus AAA91328 has been mutated more than others and is most dangerous for humans because the previous ones are similar to each other. The more left is the node from which proteins derive, the greater is the difference between them.

As it can be seen, the proteins of virus number ACI62068, isolated in Finland, and virus number ACI94923, isolated in Slovenia, are on the same scale level, which means that they have 99% sequence coverage. Thus scientists are able to divide into groups virus proteins and to investigate to determine the various mutations of viruses.

The circular phylogenetic tree consisting of all existing isolates of neuraminidase of influenza virus, isolated in human in Europe is shown in *figure 5*.

4. Conclusion

The results outcoming from this research can be formulated as follows. Highly variable nucleotide sequences in different isolates of the influenza A virus that are homologous to host genome sequences have been founf. This means that the influenza virus exchanges genetic information with the host, most probably via homologous RNA-RNA recombination. Due to this fact, the influenza virus genome besides point mutations (coming from the imprecise work of the RNA polymerase) also contains block mutations. This finding would explain the reason for the extremely fast changeability, evolution and increased virulence of the influenza A virus and its easy adaptation towards new hosts. The results outcoming from the research

will be used for development of virtual models describing the mechanism of influenza virus interaction with the host genome. This model will be applied for forecasting of appearance of new highly virulent and adaptive viral strains as a function of the host genome specificity. It will be applied also for estimation of the probability for occurrence of new epidemics and pandemics. The future work includes also computer modeling of the 3D structures of flu viral proteins (H and N), which are responsible for its virulence and propagation. On the other handthe latter model will be used for computer simulations of protein interactions with popular antiviral drugs such as Tamiflu, Relenza, Flumadin, etc. and the influence of selected gene mutations on these interactions. It is expected that the new models will find application also in the in silico drug design of new antiviral drugs and vaccines specific for the influenza A virus.

Acknowledgement

This work was financially supported by the PRACE 2 IP, WP3 (Dissemination), funded in part by the EUs 7th Framework Program (FP7 2007-2013) under grant agreement no. RI-211528 and FP7-261557. The work is achieved using the PRACE Research Infrastructure resources IBM Blue Gene/P computer located in Sofia, Bulgaria.

Manuscript received on 19.11.2012

References

1. Altschul, S. et al. Basic Local Alignment Search Tool. – J*ournal of Molecular Biology*, 215(3), 1990.

2. Altschul, S. et al. Gapped BLAST and PSIBLAST: a New Generation of Protein Database Search Programs. – *Nucleic Acids Research*, 25, 1997, 3389-3402.

3. Darling, A., L. Carey and W. Feng. The Design, Implementation, and Evaluation of mpiBLAST. Proceedings of the Cluster World Conference and Expo, in Conjunction with the 4th International Conference on Linux Clusters: The HPC Revolution, 2003.

4. Thompson, J., D. Higgins, T. Gibson. ClustalW: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. – *Nucleic Acids Research*, 22, No. 22, 1994, 4673-4680.

5. GenBank. http://www.ncbi.nlm.nih.gov/Genbank/

6. Unipro UGENE: Integrated Bioinformatics Tools.

http://ugene.unipro.ru/.

7. Etherington, G., J. Dicks, I. Roberts. Recombination Analysis Tool (RAT): a Rrogram for the High-Throughput Detection of Recombination. – *Bioinformatics*, 21 (3), 2005, 278-281.

8. Penny, D., M. Hendy, M. Steel. Progress with Methods for Constructing Evolutionary Trees. – *Trends in Ecology and Evolution*, 7, 1992, 73-79.

Plamenka Borovska
Contacts:
Technical University of Sofia, Computer Systems Department
e-mail: pborovska@tu-sofia.bg

Veska Gancheva
Contacts:
Technical University of Sofia, Computer Systems Department
e-mail: *vgan@tu-sofia.bg*

Emilyan Asenov
Contacts:
Technical University of Sofia, Computer Systems Department
e-mail: *emilyan.asenov@gmail.com*

Ivailo Georgiev
Contacts:
Technical University of Sofia, Computer Systems Department
e-mail: *ivailo georgiev@tu-sofia.bg*