Simulation of Influenza Pandemic Based on Genetic Algorithm and Agent-Based Modeling: A Multi-objective Optimization Problem Solving

Ria Lestari Moedomo1), Adi Pancoro2), Jorga Ibrahim1), Adang Suwandi Ahmad1), Muhammad Sukrisno Mardiyan1), Mohammad Bahrelfi Belatiff2), and Hengki Tasman3)

1) School of Electrical Engineering and Informatics, Bandung Institute of Technology
2) School of Life Science and Technology, Bandung Institute of Technology
3) Faculty of Mathematics and Natural Sciences, Bandung Institute of Technology
e-mail: rialmoedomo@students.itb.ac.id

Received 30 July 2009, accepted for publication 27 August 2010

Abstract
This paper describes the analysis, design and development process of simulation software for the Avian Influenza (H5N1) viruses mutation. Influenza Pandemics, which have occurred since 1729, caused by mutation (antigenic drift) and recombination (antigenic shift) of Influenza viruses. The purpose of this research is to define the modeling of virus mutation causing the Influenza Pandemic phenomena. Additionally, the objective of this simulation is to obtain all possible virus strains might be formed from mutation, the scope within this article, which can potentially trigger Influenza Pandemic. These new strains could then be utilized to support the vaccine planning process.

The Influenza Pandemic simulation program can be developed based on Genetic Algorithm method, for solving this multi-objective optimization problem. By utilizing the Genetic Algorithm approach, the chromosome solution and fitness values/functions of Influenza Pandemic stages are defined and the maximum fitness values are to be searched. The simulation result of H5N1 mutation gave 3 (three) best fitness values and a more dynamic mean fitness values, including best fitness value from several mutations combination. Simulation program was developed by utilizing MATLAB© software, with Genetic Algorithm Toolbox provided.

Keywords: Avian Influenza, Influenza Pandemic, Mutation, Multi-objective Optimization, Genetic Algorithm, Agent-based modeling

1. Introduction
In this world, Influenza pandemic has occurred several times and has caused high number of death in human population. Spanish Flu Pandemic in 1918, which caused the death of 40 million people in North America and Europe, was caused by H1N1 mutation process. Other pandemic, Asian Flu pandemic in 1957, Hongkong Flu pandemic in 1968, and Russian Flu pandemic in 1976 were caused by recombination process (Chai, 2005).

Recombination (Antigenic Shift), a sudden change of RNA viruses, enabling two different Influenza virus subtypes/strains combined or amalgamated to become a new virus subtype. On the other hand, Mutation (Antigenic Drift), a virus change/mutation process, occurs in a longer period of time. The mutation can happen within the 8 segments of Influenza virus A: HA (hemagglutinin); NA (neuraminidase); NP (nucleoprotein); M (matrix protein); PB1; PB2; PA (polymerase); and NS (non structural proteins) segments, with total 890 – 2341
nucleotide bases, can be represented in ‘G’ (guanine), ‘C’ (cytosine), ‘A’ (adenine) and ‘T’ (thymine) format. Mutation study mainly focuses on mutation on antigen surface of HA and NA segments. New virus strains formed from mutation then replaced the previous virus strain, which caused the existing vaccine to be no longer effective, hence new vaccine production is required. It was also identified that mutation can cause endemic (AI spread among avian/animals occurred in a certain region/area) which happen periodically every 1-3 years.

The Avian Influenza pandemic possibilities can be described in Figure 1 (Chai, 2005), where (B) and (D) are the mutation process:

(B) Certain Avian Influenza virus (which can only live in avian with α2-3 receptors) infects pigs (animal with α2-6 and α2-3 receptors). Then in pig’s body this virus mutates to become a new virus type/strain, which can infect other mammals and human, with α2-6 receptors.

(D) Certain Avian Influenza virus (which can live in mammals with α2-3 receptors, and in some cases live in mammals and human with α2-6 receptors) can infect human directly (with α2-6 receptors). This virus mutates to become a new virus (without any intermediary), and can easily infect other mammals and human, with α2-6 receptors.

Figure 1. Influenza Pandemic Potential through Mutations (Antigenic Drift) (Chai,2005)
Another approach was described in Shoham (2006), which stated that pandemic process can be triggered by three stages of virus change, 

(I) Increment of virus’ infectivity ability/level, from avian/animal to human (this process occurs in avian/animal body) 

(II) Increment of virus’ virulence level from low pathogenic to high pathogenic (this process can occur in either avian/animal body or human body) 

(III) Increment of virus’ contagiousness ability/level between human (this process occurs in human body) 

Each of stage/sub-process above can be mutation (antigenic drift); recombination and re-arrangement (antigenic shift). Possible process flows are: (1) Sub-process (I) only; (2) Sub-process (I) followed by (II); (3) Sub-process (I) and (II) occur in parallel, followed by (III); (4) Sub-process (I), (II) and (III) occur in sequence; (5) Sub-process (I) followed by (II) and (III) in parallel; (6) Sub-process (I), (II) and (III) all in parallel (Chai, 2005).

In this article, we focus our attention to the sub-process (I): the infectivity level increment. From Yamada (2006) which describes the mutation effects to receptor binding specificity/infectivity level, we interpret the effect of several mutations to the infectivity level increment into Table 1. Mutations and their interpreted infectivity level fluctuations are listed in the following Table 1.

**Table 1. Interpretation of several mutations effect to infectivity level increment.**

<table>
<thead>
<tr>
<th>Sialyglycopolymer Assay</th>
<th>Infectivity level increment is represented as mean value of absorbance against highest absorbance value possible</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α 2-6</td>
</tr>
<tr>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td>0.010</td>
<td>0</td>
</tr>
<tr>
<td>0.039</td>
<td>0.05</td>
</tr>
<tr>
<td>0.156</td>
<td>0.04</td>
</tr>
<tr>
<td>0.625</td>
<td>0.06</td>
</tr>
<tr>
<td>2.500</td>
<td>0.08</td>
</tr>
<tr>
<td>10.000</td>
<td>0.1</td>
</tr>
<tr>
<td>Mean value for assay = 1.667</td>
<td><strong>0.28/8 = 0.038</strong></td>
</tr>
</tbody>
</table>

**Table 1b.** Ctld (Q192R) mutation

<table>
<thead>
<tr>
<th>Sialyglycopolymer Assay</th>
<th>Infectivity level increment is represented as mean value of absorbance against highest absorbance value possible</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α 2-6</td>
</tr>
<tr>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td>0.010</td>
<td>0</td>
</tr>
<tr>
<td>0.039</td>
<td>0.25</td>
</tr>
<tr>
<td>0.156</td>
<td>0.4</td>
</tr>
<tr>
<td>0.625</td>
<td>0.6</td>
</tr>
<tr>
<td>2.500</td>
<td>0.7</td>
</tr>
<tr>
<td>10.000</td>
<td>0.7</td>
</tr>
<tr>
<td>Mean value for assay = 1.667</td>
<td><strong>2.6/8 = 0.325</strong></td>
</tr>
</tbody>
</table>

**Table 1c.** Ctld (N193K) mutation

<table>
<thead>
<tr>
<th>Sialyglycopolymer Assay</th>
<th>Infectivity level increment is represented as mean value of absorbance against highest absorbance value possible</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>α 2-6</td>
</tr>
<tr>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td>0.010</td>
<td>0.05</td>
</tr>
<tr>
<td>0.039</td>
<td>0.05</td>
</tr>
<tr>
<td>0.156</td>
<td>0.1</td>
</tr>
<tr>
<td>0.625</td>
<td>0.2</td>
</tr>
<tr>
<td>2.500</td>
<td>0.3</td>
</tr>
<tr>
<td>10.000</td>
<td>0.4</td>
</tr>
<tr>
<td>Mean value for assay = 1.667</td>
<td><strong>1.0/8 = 0.131</strong></td>
</tr>
</tbody>
</table>

Mutation process can occur within small scale or large scale (Voyles, 2002). Small scale mutation includes: 

A. **Point mutation:** the most frequent mutation is nucleotide transition: A ↔ G, C ↔ T, less frequent ones are C/T ↔ A/G transversion. 

B. **Insertion:** adding one or several nucleotides into DNA/RNA. 

C. **Deletion:** deleting one or several nucleotides from DNA/RNA. 

Besides the small scale mutation above, larger scale mutation includes: **Amplification, Deletion and Chromosomal translocation.** In this article, the scope of research is restricted to small scale mutation. 

Beside that, the mutation process is driven by other parameters, such as: only certain base positions which undergone mutation, with their specific impacts; certain host types; and the possibility of successive mutation occurrences. 

The scope of this article is restricted to the analysis of mutation of Influenza virus, A type, H5N1 subtype. The discussion was focused on HA segment of H5N1, which undergone most frequent mutations affecting the infectivity level increment.
The benefit of computation as supporting tools in Biology field has been widely known and accepted. In Bioinformatics field, the genetic mutation analysis process and the phylogenetic tree building has been done by multiple alignment and tree building software tools, for example those provided by NCBI (National Center for Biotechnology Information) in its website [http://www.ncbi.nlm.nih.gov]. With computer tools, several Bioinformatics researches can be simplified, so that the research process can be done more accurately, efficiently, and effectively.

As described above, the Influenza pandemic is triggered by some possible circumstances, as a combination of 3 sub-processes with 6 possible process flows, can be categorized as multi-objective optimization problem. Multi-objective optimization problem can be solved by genetic algorithm approach/objectives, described in the following section.

2. Methods
2.1 Genetic Algorithm Principle

The Pandemic simulation program was developed based on Genetic Algorithm principle, which is a searching algorithm based on natural selection on genetics.

In Genetic Algorithm, several aspects are defined:

- **Representation of the genetic solution domain**, known as chromosome solution, which was the solution definition with several key parameters, represented by array of bits.
- **Fitness values/function** to evaluate the solution domain. The fitness values are obtained from the fitness function, supported by statistical data analyzed from chromosome existing population.

This algorithm combines best values from chromosome community and their changes to form an innovated searching algorithm (Goldberg, 1989).

As part of genetic algorithm, there are some processes such as selection, to select some chromosome with highest fitness values as parents; reproduction; crossover; and mutation of chosen from the selection process.

In general, the genetic algorithm process flow is described in Figure 2:
As depicted in the Figure 2, the selection process in chromosome population is conducted first, then the reproduction, crossover and mutation processes are conducted in parallel. These subprocesses were done iteratively until the termination criterion is achieved. Genetic algorithm can also solve the multi-objective optimization problems, as described in the following section.

2.2 Multi-objective Optimization (MOP)

Most realistic optimization problems, particularly those in design, require the simultaneous optimization of more than one objective function. Multi-criteria optimization has its roots in late-nineteenth-century welfare economics, in the works of Edgeworth and Pareto (Das, 1997). The objective of MOP is to obtain the optimum values from set of objective functions, with a mathematical description as follows:

\[
\min F(x) = \begin{bmatrix}
    f_1(x) \\
    f_2(x) \\
    f_3(x) \\
    \vdots \\
    f_n(x)
\end{bmatrix} \quad \text{(MOP)}
\]

where \( x \in C \) and \( n \geq 2 \) and

\[
C = \{ x : h(x) = 0, g(x) \leq 0, a \leq x \leq b \}
\]

denotes the feasible set constrained by equality and inequality constraints and explicit variable bounds. The space in which the objective vector belongs is called the \emph{objective space} and image of the feasible set under \( F \) is called the \emph{attained set}.

The scalar concept of "optimality" does not apply directly in the multi-objective setting. A useful replacement is the notion of \emph{Pareto optimality}. Essentially, a vector \( x^* \in C \) is said to be Pareto optimal for MOP if all other vectors \( x \in C \) have a higher value for at least one of the objective functions \( f_i(\ast) \), or else have the same value for all objectives.

The multi-objective problem can almost always be solved by combining the multiple objectives into one scalar objective whose solution is a Pareto optimal point for the original MOP. Most algorithms have been developed in the linear framework (i.e. linear objectives and linear constraints), but the following technique is applicable to nonlinear problems.

A standard technique for MOP is to minimize a positively weighted convex sum of the objectives, that is,

\[
\sum_{i=1}^{n} \alpha_i f_i(x), \alpha_i \neq 0, i = 1, 2, 3, \ldots, n
\]

It is easy to prove that the minimizer of this combined function is Pareto optimal. It is up to the user to choose appropriate weights. Until recently, considerations of computational expense forced users to restrict themselves to performing only one such minimization. Newer and more ambitious approaches aim to minimize convex sums of the objectives for various settings of the convex weights, therefore generating various points in the Pareto set. This multi-objective equation will be implemented for modeling the Pandemic phenomena, which was the scope of this research.

2.3 Agent-based modeling

Agent based modeling (ABM) is a stochastic modeling used to describe populations of interacting agents, such as insects and people, using simple rules that dictate their behaviors (Bauer, 2008). ABM has benefits compared with other modeling techniques, i.e. (i) ABM captures emergent phenomena; (ii) ABM provides a natural description of a system; and (iii) ABM is flexible. However, the ability of ABM to deal with emergent phenomena is what drives the other benefits (Bonabeau, 2002). Related to interaction tasks, Fachada’s research (Fachada, 2008) described how to utilize ABM to model the Avian Influenza antigenic variability and the infected human immune system while Bauer’s research described how ABM was utilized to model general and specific immune systems and disease simulators (Bauer, 2008).

For the Avian Influenza phenomena, ABM can represent the interactions between the Avian Influenza viruses (agents) and their hosts (agents), and their changes because of these interactions. The important role of interaction between mutated viruses and their receptors are clearly described in the new-strike research article (Chandrasekaran, 2007). This new study suggests that the difference in binding preference between human and avian flu viruses is more complicated than just \( \alpha_{2-6} \) versus \( \alpha_{2-3} \). Rather, it is an affinity for a particular topology, or shape, of \( \alpha_{2-6} \) glycan receptor that characterizes human flu viruses. Specifically, the human flu viruses prefer long \( \alpha_{2-6} \) receptors that occupy an umbrella-shaped space, as opposed to \( \alpha_{2-3} \) and some \( \alpha_{2-6} \) receptors that occupy a cone-shaped space (Ross, 2008).
year 2000 to 2008. The query result was 231 nucleotide sequences, and these virus sequences were utilized for mutation analysis task, to be described later on in section 2.4.2.

Chromosome solution was represented as composition of several information, that is:

1. HA virus sequence, obtained from virus sequences database, in this case NCBI. There are 1870 bases, each base has 4 possible values (‘G’, ‘C’, ‘A’, ‘T’), and can be represented into 2 bits. Total bits required to represent 1 virus sequence was $2 \times 1870$ bits = 3740 bits. Sample of HA sequence is depicted in Appendix A.

2. Host type, several possible hosts includes: avian, blow fly, camel, canine, cat, civet, equine, giant anteater, human, leopard, mink, mouse, seal, stone marten, swine, tiger and whale, which sum up to 17 hosts. For H5N1 population in Indonesia, there are only four possible hosts: avian, human, cat, and swine. Host types can be represented by 2 bits ($2^2 = 4$ possibilities).

2.4.2 Pandemic process modeling

As described in the Introduction section, Pandemic process can be triggered by 3 stages of virus changes, with 6 possible flow possibilities as in Figure 3. These sub-processes were modeled as Multi-Objective Optimization Problem. In process flow (4) in Figure 3 (which will be the focus of this article), each sequential sub-process is represented by an objection function: (I) represented by Objective Function I, (II) represented by Objective Function II, (III) is represented by Objective Function III.

![Diagram](image)

**Figure 2.** Several possible Influenza pandemic process-flows

2.4.2.1 Infectivity Objective Function

This function is related to infectivity optimization, aims to obtain virus strain with highest infectivity level. Determined sub-functions of Objection Function I are:

$$ F_1 = (F_{1a} \cdot b_{1a}) + (F_{1b} \cdot b_{1b}), $$
with:

- $F_{1a}$ is the internal sub-function which represents the occurrence of a certain mutation, which caused the infectivity increment. Determined parameters are: mutation occurrence, infectivity level, infectivity increment, its weighted factor, host probability and probability of certain mutation occurrence:

$$F_{1a} = \text{infectivity} \times 100\% + \sum (\text{mutation occur} \times \text{infectivity incr} \times \text{mutation weight} \times \text{mutation prob})$$

- $\text{infectivity}$: absorbance value (measurement: nm) obtained from Biomolecular Lab's measurement
- $\text{mutation occur}$: indication of occurrence of certain mutation (value: 0/1)
- $\text{infectivity incr}$: increment of infectivity level caused by a certain mutation (measurement: %) which is the result from graph in Yamada (2006)
- $\text{mutation weight}$: certain mutation weight based on interpretation of mutation impact to infectivity increment

$\text{mutation weight}$ and $\text{mutation prob}$ variables above are required to obtain the relative/estimated infectivity value of mutation which infectivity increment is not known. For example, from Table 2, if the infectivity increment value ($I$) from base mutation 63-65 AAT $\rightarrow$ GAT is known, to estimate the infectivity increment value ($I'$) of base mutation 63-65 AAT $\rightarrow$ AAC, we define the following equation: $I' = I \times 20/33.8$.

- $b_{1a}$: weight of internal infectivity sub-function relative to external infectivity sub-function

- $F_{1b}$ is the external sub-function based on mean value of certain virus occurrence in several other hosts in the same location:

$$F_{1b} = \sum ((\text{virus freq in other host} / \text{virus freq total}) + \text{host prob}/2) \times \text{host impact scale}) / \text{number of hosts}$$

- $\text{virus freq other host}$ and $\text{virus freq total}$: local variable which is (1) number of cases of a certain virus strain occurrence on other hosts and (2) number of cases of a certain virus strain occurrence on all hosts.
- $\text{host prob}$: this global variable is the probability value of a certain host getting infected by a certain virus strain observed from statistical analysis NCBI database
- $\text{host impact scale}$: impact scale of a certain host relative to other hosts
- $\text{number of hosts}$: number of hosts within a certain location

- $b_{1b}$: weight of external infectivity sub-function relative to internal infectivity sub-function

To obtain the statistics for Infectivity level measurement, the secondary data analysis was done by observing all the existing H5N1 virus sequences, as in Table 2, from the Avian Influenza virus database and tools provided in NCBI website.

Table 2 describes the base position/amin acid of virus strain undergone mutation; mutation impact of such mutation; pandemic sub-process triggered by this mutation; infectivity level increment caused by this mutation (the values are obtained from Table 1); and weighted factor of each mutation.

<table>
<thead>
<tr>
<th>No</th>
<th>Base position</th>
<th>Mutation impact</th>
<th>Pandemic stage/sub-process</th>
<th>Infectivity Level Increment</th>
<th>Weighted Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>544-546 (amino acid position: 182), mutation from Asn $\rightarrow$ Lys</td>
<td>Can be bound to both avian and human receptors</td>
<td>I</td>
<td>5%</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>574-576 (amino acid position: 192), mutation from Gln $\rightarrow$ Arg</td>
<td>Can be bound to both avian and human receptors</td>
<td>I</td>
<td>46%</td>
<td>1</td>
</tr>
<tr>
<td>3.</td>
<td>577-579 (amino acid: 193), mutation from Asn $\rightarrow$ Lys</td>
<td>Can be bound to both avian and human receptors</td>
<td>I</td>
<td>33%</td>
<td>1</td>
</tr>
<tr>
<td>4.</td>
<td>667-669 (amino acid position: 223) of swine influenza virus</td>
<td>Increases virus ability to bind to human receptors, and reduces its affinity for poultry receptors</td>
<td>I</td>
<td>To be estimated</td>
<td>0.7</td>
</tr>
<tr>
<td>5.</td>
<td>676-678 (amino acid position: 226)</td>
<td>Reduce binding affinity to poultry receptor</td>
<td>I</td>
<td>To be estimated</td>
<td>0.7</td>
</tr>
<tr>
<td>6.</td>
<td>679-681 (amino acid position: 227)</td>
<td>Reduce binding affinity to poultry receptor, increase binding affinity to human receptors</td>
<td>I</td>
<td>To be estimated</td>
<td>0.7</td>
</tr>
<tr>
<td>7.</td>
<td>682-684 (amino acid position: 228)</td>
<td>Reduce binding affinity to poultry receptor</td>
<td>I</td>
<td>To be estimated</td>
<td>0.7</td>
</tr>
</tbody>
</table>
All mutations occurred were analyzed by the multiple alignment process, which compares all HA sequence with the consensus sequence. The HA sequence used as consensus was AF144305, which is one of the ancestors of H5N1 viruses in Asia. The multiple alignment process was executed to obtain information of all mutation occurred in each virus’ sequence bases, as in Figure 4. With the use of consensus AF144305 HA sequence (the most upper row), the mutation types occurred in all 231 HA sequences are compared and analyzed, including: transition (A $\rightarrow$ G, C $\rightarrow$ T), transversion (C/T $\leftrightarrow$ A/G) and gap - deletion/insertion (‘-’). As an example, from sequence AM18366,A/chicken/Indonesia /R134/03 (H5N1), gaps occurred between position 6-12, transversion A $\rightarrow$ C in position 21, transition C $\rightarrow$ T in position 78, 120, transition G $\rightarrow$ A in position 123, etc, until the analysis reaches the end of the sequence: position 1870. The same analysis process should be conducted for all nucleotide sequences, however in this article the analysis was done only for base position 1 to 70 from the 231 nucleotide sequences.

The result were statistics of several key parameters of mutation: mutation impact, host types and mutation probabilities as in Table 2, 3 and 4. Each of these supporting statistics is described in the following paragraph:

![Figure 4. Result of HA H5N1 sequences multiple-alignment](image)

(1) Mutation impact of certain mutated bases

From all mutation occurred, the impact analysis was recorded, to identify the virus characteristics changes cause by mutation of certain bases’ position. As an example: In H5N1 virus, the HA change in amino acid (bases triplet) in position 182 and 192 can enable virus to be bound to both avian and human receptors (Yamada, 2006).

The definition process of mutation impact was done, by studying H5N1 mutation references, and by studying and analyzing the H5N1 amino acid mutation motives and their impacts.

From mutation occurrence, we can determine whether or not such mutation has bigger influence towards Pandemic possibility. If it has a bigger influence, then it is given higher fitness value. For example, if the mutation causes the virus bounded to several hosts, then the fitness value is bigger than other mutations which cause the virus has the ability to change its host, but lost its binding ability to the previous host.

We can also conclude if such mutation occurs in pandemic sub-processes (I), (II) or (III), to be determined its weighted factor. As an example, sub-process (I), which was the increment of virus infectivity ability from avian/animals $\rightarrow$ human was given higher weighted factor compared to the (II) and (III) sub-processes, because its greater impact toward the pandemic occurrence, and also because all possible pandemic process flows always begin with sub-process (I). In this case, the weighted factor for sub-process (I) was determined as 1, sub-process (II) and (III) was determined as 0.5, as in Table 2.

(2) Host Type

There are only four host types of Avian Influenza occurred in Indonesia, with statistics as in Table 3.
Table 3. H5N1 statistic according to host types

<table>
<thead>
<tr>
<th>No.</th>
<th>Host types</th>
<th>Occurrence</th>
<th>(Fitness Value)</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Avian</td>
<td>140</td>
<td></td>
<td>60.7</td>
</tr>
<tr>
<td>2.</td>
<td>Human</td>
<td>88</td>
<td></td>
<td>38.1</td>
</tr>
<tr>
<td>3.</td>
<td>Cat/Other mammals</td>
<td>1</td>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td>4.</td>
<td>Swine</td>
<td>2</td>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>231</td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

(3) Mutation Types and Probabilities

With the use of consensus AF144305 HA sequence (the most upper row), the mutation types occurred in all 231 HA sequences are compared and analyzed, including: transition (A↔G, C↔T), transversion (C/T ↔ A/G) and gap/deletion/insertion. At the end, statistic and fitness values for all 231 nucleotides are obtained. As an example, from the analysis of position 1 to 70 from the 231 nucleotide sequences, we obtain the mutation statistic samples in Table 4:

Table 4. H5N1 strains position 1-68 statistical analysis according to mutation types

<table>
<thead>
<tr>
<th>No.</th>
<th>Base position</th>
<th>Mutation type</th>
<th>Occurrence (Fitness value)</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>15</td>
<td>Transversion A → C</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td>2.</td>
<td>21</td>
<td>Transversion A → C</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>3.</td>
<td>22</td>
<td>Transition T → C</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>4.</td>
<td>40</td>
<td>Transition A → G</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>Transversion A → C</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subtotal</td>
<td></td>
<td>4.5%</td>
</tr>
<tr>
<td>5.</td>
<td>45</td>
<td>Transition A → G</td>
<td>5</td>
<td>7.7%</td>
</tr>
<tr>
<td>6.</td>
<td>46</td>
<td>Transition A → G</td>
<td>3</td>
<td>4.6%</td>
</tr>
<tr>
<td>7.</td>
<td>48</td>
<td>Transition A → G</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td>8.</td>
<td>50</td>
<td>Transition T → C</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>9.</td>
<td>54</td>
<td>Transition T → C</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>10.</td>
<td>59</td>
<td>Transition T → C</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td>11.</td>
<td>60</td>
<td>Transition T → C</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td>12.</td>
<td>63</td>
<td>Transition A → G</td>
<td>22</td>
<td>33.8%</td>
</tr>
<tr>
<td>13.</td>
<td>64</td>
<td>Transition A → G</td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>Transversion A → C</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subtotal</td>
<td></td>
<td>4.5%</td>
</tr>
<tr>
<td>14.</td>
<td>65</td>
<td>Transition T → C</td>
<td>1</td>
<td>20%</td>
</tr>
<tr>
<td>15.</td>
<td>68</td>
<td>Transition A → G</td>
<td>2</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>Transition T → C</td>
<td></td>
<td>1</td>
<td>1.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subtotal</td>
<td></td>
<td>4.5%</td>
</tr>
</tbody>
</table>

This process should be done for all mutation types occurred in all 231 nucleotide sequences from position 1-1870. Therefore, it is necessary to build some supporting software tools for this analysis process automation purposes.

(4) Host Impact Scale

This scale was required in Function F1b, to determine the Infectivity level of a virus strain. Each host was given a value, to determine its impact towards the Pandemic possibility, if there was any
virus strain within such host. In Figure 5, the host with highest impact scale are human and swine:

1: Whale
3: Avian
5: Swine
7: Human

Figure 5. Host Impact Scale

2.4.2.2 Virulence Objective Function

This function is related to Virulence Optimization, aims to obtain virus strain with highest Virulence level. To evaluate the virulence and pathogenic, there are several parameters to be considered (Conenello, 2007): (1) Morbidity (decrement of weight host); (2) Titer LD<sub>50</sub> (number of virus in host’s cell body); (3) Virus replication. For this research, only Morbidity is taken into consideration.

2.4.2.3 Contagiousness Objective Function

This function is related to Contagiousness Optimization, aims to obtain virus strain with highest Contagiousness level. Contagiousness level increment is represented as secondary attack rate function (Eichner, 2009); (Cannell, 2008), which is the division between secondary cases (number of population infected from primary cases, which is number of population infected in the first time period) and number of population within a particular community vulnerable to influenza virus infection.

Certain weight factors were already incorporated within the functions. For example, the weight factors were determined as:
- Weight factor for Function I = 35%
- Weight factor for Function II = 20%
- Weight factor for Function III = 45%

To solve the Pandemic Multi-Objective Optimization problem, the Pareto Optimality of new virus strains generated by Genetic Algorithm, resulting from Function I, II and III should be conducted. However the scope of this article is only Infectivity Objective Function (Objective Function I).

2.5 Virus–Host Interaction Modeling

To define the interaction between viruses and their hosts, the attributes (properties) and behavior of each agent need to be described. For example, attributes belong to viruses include: type of virus (H5N1, H1N1, H2N1, H7N7); virus sequence: HA sequence and NA sequence; Infectivity Level; Virulence Level; Contagiousness Level and Mutations occurred within a specific virus, while attributes belong to hosts include: Host species (Avian such native chicken, broiler, duck; Human; Swine; Etc); Receptors types (α<sub>2</sub>-3 or α<sub>2</sub>-6) and its shape: corn-like or umbrella-like; and other host health condition parameters. This agents modeling can be designed by Agent-based modeling tools such as REPAST (Etatara, 2007), as depicted in Figure 6 (Native chicken host model), and the interaction between agents was depicted in behavior script written in REPAST language.

Figure 6. Chicken host agent modeling (Etatara, 2007)

2.6 Validation Method

The simulation program would generate a number of possible mutated virus strains with their fitness values. The simulation should produce both the newly generated virus strains and their fitness values. Based on the highest fitness values, several virus strains are chosen and determined as virus strains which could potentially cause Influenza Pandemic. To validate these virus strains, it is required to obtain real virus strains from certain period of time in the past including all mutated strain viruses and their infectivity, virulence and contagiousness values, to be compared with the result from this experiment. For example: (1) H5N1 virus strains from the beginning of endemic in Indonesia around 2003, and (2) H5N1 virus strains from year of 2008 and real infectivity level values from Lab’s measurement, to be compared with result from this experiment. If the infectivity level values obtained from this experiment are closely similar to infectivity level from Lab’s measurement, then this experiment is proven correct. However, until now the real infectivity level values from Lab’s measurement are not yet obtained.
3. Result and Discussion

To develop Pandemic Simulation prototype, MATLAB® software with Genetic Algorithm toolbox was utilized.

The process is as follows:
(1) Defining the chromosome: strain virus (1680 base x 2 bits) and its host (2 bits). In this research, we focus only on HA base position 541-550 and 571-580 because these two portion of virus stain undergone mutation. An example of virus strain input is as follows:
Strain virus H5N1 HA base position 541-550 and 571-580 as follows:

GGGAAAGGGG  GGGCGAGGGGG
541-550               571-580

With convention: G = 00 ; C = 01 ; A = 10 ; T = 11,
this strain virus on avian is converted to :

0000001010100000000000000000000000000000000000 00
540              570          Host type

This string is then converted to decimal format, for simulation program input:

4.509717504000000e+010

(2) Defining the fitness function for infectivity optimization, as described in the previous section.
(3) Performing the Genetic Algorithm simulation, to obtain all possible viruses strain, and choose those with highest fitness values.

The Objective Function was defined by MATLAB® function. There are 2 mutation defined: N182K (base position 544-546) and Q192R (base position 574-576) and the string inputs randomly obtained from GA toolbox reside both in avian and human.

On MATLAB®’s Genetic Algorithm Toolbox, the parameters are set as follows: (1) Fitness function: @infektivitas_mutasi; (2) Population size: 20; Initial population: [ ]; Initial range: [4.509717504000000e+010 ; 1.049780741169885e+012] and Cross-over fraction: 0.514286.

The result of this initial prototype as depicted in Figure 7, was 51 generation with 3 best fitness values:

(1) 4.5 from strain virus undergone N182K mutation in human host; (2) 15.8 from strain virus undergone N182K mutation in avian host; (3) 17.2675 from strain virus undergone N182K Q192K mutation in avian host.

![Figure 7. H5N1 mutation simulation result](Image)

This prototype is further developed by doing the following tasks:
(1) Completion of fitness values parameter (1) and (3). The fitness values of parameter (1) can be obtained by analyzing the H5N1 mutation
characteristics and impact references, and then determining the correct fitness function. Parameter (3) can be obtained by analyzing the H5N1 virus sequences secondary data from existing virus sequences databases, with support of software tools (to be developed), to obtain the fitness values automatically.

(2) Completion of all fitness functions definition, based on multi-objective optimization principle, for each of Influenza pandemic sub-processes.

(3) Obtaining all optimal fitness values and the corresponding newly-generated virus strains

(4) Accommodating the requirements in obtaining a more comprehensive viruses stains database, including all information related to fitness function parameters,

(5) Performing a more comprehensive exploration of MATLAB© and its Genetic Algorithm toolbox, as well as Agent-based program and integration between these two approaches.

4. Conclusions

From the analysis process, the data and process modeling for several mutation parameters has been done, with (1) genetic chromosome solution and (2) objection function definition as the result. With these two inputs, the initial prototype was developed by utilizing MATLAB© software. As in Figure 7, this simulation result gave 3 fitness values as explained in previous section. Better and more fitness values, including fitness values from several mutations combination, could be obtained once the simulation program and virus mutation database are developed and enhanced.

Beside completion of this initial prototype according to development plan described in Section 3, study about how to implement the Recombination process with MATLAB© Genetic Algorithm Toolbox is required, to provide the application platform for an overall research scope. The Agent-based modeling tool and program also has to be explored in more depth, as Repast visual-modeling-tool, although provides easiness for design purposes, did not give satisfactory result for program code generating purposes. Therefore other tool/programming language and Repast Java-based tool are required to be explored as this application platform and then integrated with current MATLAB© program.

5. Acknowledgements

This article is dedicated to the late Ibu Ietje Suwenda Saleh, former Biology teacher in SMAN 3 Bandung, who had given inspirations to start this research.

Sincere thanks are directed to Deni and Dr. Agus Yodi Gunawan from Mathematics ITB, for valuable inputs in defining the Multi-objective functions and Genetic Algorithm aspects; Dr. Bambang Riyanto and Dr. Arief Syaichu Rohman from STEI ITB who kindly gave supervision on MATLAB© software; Prof. Richard Belew, from University of California in San Diego (USCD), for valuable inputs and several papers related to this research.

References


Ross, Robert, January 2008, Study refines view of H5N1 virus's binding preferences, CIDRAP (Center for Infectious Disease Research and Policy Academic Health Center, University of Minnesota).


Appendix A.

Sample of H5N1’s HA sequence information from a virus strain in Indonesia: A/Ck/Indonesia/BL/2003

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>AY651321</th>
<th>1659 bp</th>
<th>cRNA</th>
<th>linear</th>
<th>VRL 10-JUL-2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Influenza A virus (A/Ck/Indonesia/BL/2003(H5N1)) hemagglutinin (HA) gene, partial cds.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACCESSION</td>
<td>AY651321</td>
<td>GI:50296026</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>Influenza A virus (A/Ck/Indonesia/BL/2003(H5N1)) Viruses; sRNA negative-strand viruses; Orthomyxoviridae; IInfluenzaVirus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

REFERENCE 1 (bases 1 to 1659)


TITLE Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia


PMID 15241415

REFERENCE 2 (bases 1 to 1659)


TITLE Direct Submission

JOURNAL Submitted (14-JUN-2004) Microbiology, The University of Hong Kong, HKU Pathology Building, Pokfulam, Hong Kong SAR, China

FEATURES Location/Qualifiers

source 1..1659 /organism="Influenza A virus (A/Ck/Indonesia/BL/2003(H5N1))" /mol_type="viral cRNA" /strain="A/Ck/Indonesia/BL/2003" /db_xref="taxon:284182" /country="Indonesia" /note="subtype: H5N1" /db_xref="GENBANK:AY651321" /gene="HA" /product="hemagglutinin" /product="HA" /version="AY651321.1" /gene="HA" /db_xref="GI:50296027" /db_xref="GAP4182" /country="Indonesia" /db_xref="GENBANK:AY651321.1" /gene="HA" /db_xref="GI:50296027"

ORIGIN

1 atggagaaaa tagtgcttct tcttgcaata gtcagtcttg ttaaaagtga tcagatttgc
61 attggttacc atgcaaacaa ttcaacagag caggttgaca caataatgga aaagaacgtt
121 actgttacac atgcccaaga catactggaa aagacacaca acgggaagct ctgcgatcta
181 gatggagtga agcctctaat tttaagagat tgtagtgtag ctggatggct cctcgggaac
241 ccaatgtgtg acgaattcat caatgtaccg gaatggtctt acatagtgga gaaggccaat
301 ccagccaatg acctctgtta cccagggaat ttcaacgact atgaagaact gaaacaccta
361 ttgagcagaa taaaccattt tgagaaaatt cagatcatcc ccaaaagttc ttggtccgat
421 catgaagcct catcaggggt gagctcagca tgtccatacc agggaaagtc ctcctttttt
481 agaaatgtgg tatggcttat caaaaagaac agtgcatacc caacaataaa gagaagctac
541 aataatacca accaagaaga tcttttggta ctgtggggga ttcaccatcc taatgatgcg
601 gcagagcaga caaggctata tcaaaaccca accacctata tttccgttgg gacatcaaca
661 ctaaaccaga gattggtacc aaaaatagct actagatcca aagtaaacgg gcaaagtgga
721 aggatggaat tcttctggac aattttaaaa ccgaatgatg caatcaactt cgagagtaat
781 ggaaatttca ttgctccaga atatgcatac aaaattgtca agaaagggga ctcagcaatt
841 atgaaaagtg aattggaata tggtaactgc aacaccaagt gtcaaactcc aatgggggcg
901 ataaactcta gtatgccatt ccacaacata caccctctca ccatcgggga atgccccaaa
961 tatgtgaaat caaacagatt agtccttgcg actgggctca gaaatagccc tcaaagagag
1021 agaagaagaa aaaagagagg actatttgga gctatagcag gttttataga gggaggatgg
1081 cagggaatgg tagatggttg gtatgggtac caccatagca atgagcaggg gagtgggtac
1141 gctgcagaca aagaatccac tcaaaaggca atagatggag tcaccaataa ggtcaactcg
1201 atcattgaca aaatgaacac tcagtttgag gccgttggaa gggaatttaa taacttagaa
1261 aggagaatag aagtaacgac ccacctcata cccggagaat cccggagaat
1321 aggagaatag aagtaacgac ccacctcata cccggagaat cccggagaat
1381 gtagaagt gccatagt oneattccc aaccccacccc ccccacccc ccccacccc
1441 gtagaagt gccatagt oneattccc aaccccacccc ccccacccc ccccacccc
1501 gtagaagt gccatagt oneattccc aaccccacccc ccccacccc ccccacccc
1561 gtagaagt gccatagt oneattccc aaccccacccc ccccacccc ccccacccc
1621 atcattgaca aaatgaacac tcagtttgag gccgttggaa gggaatttaa taacttagaa