Preface
The following document is the first version of the system requirement for the ZooPhy interface project. The project customer contact is Dr. Matthew Scotch (email: Matthew.Scotch@asu.edu). The following personnel are assigned to the project:
- Gazi Islam
- Pierina Ortiz
- Ashutosh Singraur

The following sections can be found in the requirement:
- Introduction: Brief introduction to project at hand, and user requirements.
- Glossary: Technical words defined.
- User requirements definition: Requirements specified by the project contact. System Architecture: Anticipated system Architecture
- System requirement specification: System components specified.
- System models: Graphical system models
- System Evolution: Anticipated changes to the system
- Risk Analysis
- Project Plan (preliminary)
- Test Plan (preliminary)
- Gantt Chart
- Appendix

Introduction
ZooPhy is a system for performing phylogeography of zoonotic RNA viruses. The end product will take input from a user interface, query the ZooPhy database, process a workflow of molecular evolution steps, and produce an output that shows the evolutionary migration of the virus within an animal host over a specific landscape (a phylogeographic tree). The users of the system will include anyone interested in surveillance of zoonotic diseases in either animals or humans as host.

The specific task at hand is to create a user-friendly interface that allows users interested in phylogeography analysis of RNA viruses, to access the database content and conduct respective analysis. The user can select a number of RNA virus, subtype and host. Other options are the date of isolation and the geographical location of the isolate. Thereafter, a table of sequences is created, which can further be used to plot a phylogeographic tree, Bayesian skyline plot or perform multiple sequence alignment of the virus.

The interface needs to allow the user to specify different attributes such as: genes, viruses, animal species, geographic area, and time-frame. For example, typical queries might be:
2. Return sequences for "Glycoprotein (G) gene, of Rabies Viruses, in New England, in raccoons, from 1975 - 2000".

These attributes have many possibilities. For example, species are often represented in taxonomy in a 'tree of life' as kingdom, phylum, class, order, rankings, etc. Thus it might be useful to allow the user to search for species by typing in a name, or a point-and-click style that allows them to iterate through a 'tree-like' taxonomy. Also, the user should also be able to specify the type of analysis they want to do. For example, 'Bayesian phylogeography' vs. Neighborhood-Joining Tree vs. Maximum Likelihood, etc (or a combination).

ZooPhy is running on a Windows server and stored in SQL Server. The server is located at Arizona State University, Biomedical Informatics Department (zoophy.asu.edu). Currently the functions of ZooPhy include:

- Bayesian phylogeographic trees *Bayesian Evolutionary Analysis by Sampling Trees (BEAST)*
- Bayesian skyline plots through *BEAST*
- Multiple sequence alignment via a web-service that connects to *Clustal*.

**Glossary**

**Host Animal**: Animals in which the virus infects and multiplies. Most of the viral strain is specific set of animals. e.g. Influenza A subtype H1N1 infects host Human, Birds, Horse and Pigs.

**Maximum Likelihood**: A criterion for estimating a parameter from observed data under an explicit model. In phylogenetic analysis, the optimal tree under the maximum likelihood criterion is the tree that is the most likely to have occurred given the observed data and the assumed model of evolution.

**Neighborhood Joining**: In bioinformatics, neighbor-joining is a bottom-up clustering method used for the construction of phylogenetic trees. Usually used for trees based on DNA or protein sequence data, the algorithm requires knowledge of the distance between each pair of taxa (e.g., species or sequences) in the tree.

**Phylogeography**: Field of study concerned with the principles and processes that govern the geographic distributions of genealogical lineages, especially those within and among closely related species. Phylogeography deals with historical, phylogenetic components of the spatial distributions of gene lineages (Avise 2000)

**Public Health Informatics**: systematic application of information and computer science and technology to public health practice, research, and learning. It is one of the sub domains of health informatics.
RNA virus: A virus that has RNA (ribonucleic acid) as its genetic material. They are classified as either single-stranded RNA (ssRNA) or double-stranded RNA (dsRNA).

Taxonomy: The classification of organisms in an ordered system that indicates natural relationships.

Tetrapods: Animals with four recognized limbs.

Virus: A small infectious agent that can replicate only inside the living cells of organisms. Virus particles (known as virions) consist of two or three parts: the genetic material made from either DNA or RNA, long molecules that carry genetic information; a protein coat that protects these genes; and in some cases an envelope of lipids that surrounds the protein coat when they are outside a cell.

Virus Subtype: A subclass of virus species. For example the virus subtype for influenza virus would be H1N1. They are basically formed due to recombination of viral genome. They are responsible for causing disease.

Zoonotic Virus: Any disease that is spread from animals to people.

Zoophy: A public health informatics application for phylogeography of zoonotic RNA viruses. The application was developed to aid in the surveillance of zoonotic diseases. For more information please refer to: http://zoophy.wikispaces.asu.edu/

User Requirements Definition

Functional Requirements
The customer expressed the following requirements for the end product of the ZooPhy interface:

1. Interface should be created in Microsoft .NET platform.
2. The query page must have a systematic representation to do the query.
   a. Name of virus, subtype, and host
   b. Provided with filters like isolation date range
   c. Provided with geographical info (Google map form)
3. Taxonomic representation of Virus
4. Taxonomic representation of Host
5. A drill down connectivity between the virus strain and the host animal should be provided for the user. For example, the strain H1N1 of Influenza A virus affects has host as human, pig, bird, horse while other subtypes can have different host animals.
   a. If the user selects a specific virus with a subtype then the host specific to subtype should be displayed.
   b. If the user only selects the virus in whole, then all the animals which are host of all subtypes should be displayed.
6. User should be able to save and share results.
7. Implementation of future natural language capabilities should be incorporated into the interface.
Non-functional Requirements
The customer expressed that the following requirements for the end product of the ZooPhy interface:

- Computer space requirement will be of less than 20mb.
- The interface should be user-friendly. It should be possible for anyone with computer knowledge to navigate through ZooPhy. Users include health agencies, agriculture agencies, wildlife agencies and other researches interested in surveillance.
- Zoophy source code for the interface will be located in the ZooPhy lab computers with backup in the personal computers from each of the individuals assigned to the project.
- The server specification are as follows:
  - Dell PowerEdge R910
  - Memory - 128gb Memory, 1066MHz, Dual Ranked RDIMMs for 4 processors
  - Processor – 6 Intel Xeon E7530 1.86GHz processors, 12M cache
  - RAID – Raid 5
- ZooPhy FAQ will link to ZooPhy wikispace: [http://zoophy.wikispaces.asu.edu/](http://zoophy.wikispaces.asu.edu/)

System Architecture
The system architecture was built according to the user requirement definition. To meet the end user requirements the following components will make up the Zoophy interface

![Figure 1: System Architecture](image-url)
Case scenarios

Scenario 1
Initial Assumption:
The male user knows exactly the virus, subtype and host that he will be conducting the analysis for.

Normal:
The user opens the web browser, he then types in the ZooPhy URL: http://zoophy.med.yale.edu/web_ZooPhy/. Since the user is a new user to the system he goes ahead and registers as a new user. The information window opens where the he is asked to give information about himself, name, user name, password, institute, e-mail address and institution address. He clicks on the register button to complete the process. He can now log into the ZooPhy website.

Once logged in, user wants performs a search to select the specific type of virus he is interested in. The user clicks on virus to start the search in the taxonomic tree. A tree generates categorizing the virus in 2 divisions, DNA virus and RNA virus. He wants to the search on RNA virus, so he clicks on the RNA virus link. As soon as he clicks on the RNA virus the tree expands which sub-dividing the RNA virus into single stranded and double stranded. By clicking on single stranded, the user notices that the corresponding hierarchical taxonomic tree is displayed to narrow down the search option. He then has to make a choice between positive sense strand and negative sense strand. On selecting positive sense strand a tree of all family members appears. Each family member includes all its members like family orthomyxoviridae contains influenza virus. On proceeding with Influenza A, a connecting tree with all its subtypes is displayed. The user decides he wants to conduct the query with only H1N1, thus he selects that subtype, he sees that the virus and subtype fields are filled out.

The user moves on to select the host animal. He notices that ZooPhy provides the list of host animals also in a taxonomic tree representation aside from being able to type in the name of the host. He first plays around with the tree and he makes the observation that if a specific subtype is selected then automatically the host search will show only those animals which are infected by the specific viral subtype. The user can select either only one host animal or a number of animals together by pressing the Ctrl button. Since this is the first time using the system, the user want to explore what happens when he selects all the host option. He does this by pressing the select all checkbox on the bottom.

The user now proceeds to select the geographic location of isolation of the virus and subtype selected. He notices that a Google map in the interface, where he can zoom in and out and select geographical regions, continent, countries, provinces. The user drags the mouse to the close to the North American continent, and pins the United States.
Now the user selects the date of isolation. He is interested in a time span that goes back 10 years from today. The user notices that he can either type in the date or select from calendar form. He selects the dates that he wants.

What can go wrong:
The user could be searching for Influenza A virus, and he wants to select insects as host. This is incorrect, since insects do not host the virus. The user may want to back to the literature search or he can look in the help page to see that the reason for the host not present is because the virus does not infect the host.

The user is not an expert in phylogenetics and does not know exactly what each field means. He should then refer to the help document.

Other activities:
The search field can be saved by the user for later retrieval of the results.

System State on Completion:
The user is logged on. All the search fields are filled out; the user can now conduct the analysis.

Scenario 2
Initial Assumption:
The user has already filled out the following fields: virus: Influenza, subtype: H1N1, host: all, geographical location: United States and date of isolation: from today to ten years back.

Normal:
After the user has completed the input page, he clicks on the search button and a display page will be in a excel sheet format. The user notices that the first column has the isolate ID of the viral strain from GenBank, or Epiflu databases. The next column contains the subtype and the third column gives the passage name. The rest of the columns shows the protein name and the length of the sequence. He scrolls up and down to view the different sequences available. The user notices that there is a checkbox on each sequence, he goes ahead and selects ten sequences, and scrolls up to the top of the page. There are a few buttons that the user notices: phylogeographic tree, Bayesian skyline plot, multiple sequence alignment. The user decides that he wants to try all the analysis. First he clicks on the multiple sequence alignment the results and performs a multiple sequence alignment (MSA). The result of the MSA can be downloaded and visualized in tools like Seaview, Clustal, Bioedit. Then he goes back to the results and clicks on the option to create a phylogeographic tree. A new window opens, showing the phylogeographic tree of the selected viral subtypes over the geographical area the user selected. The user is very pleased on how the result is displayed. Thereafter, the user conducts Bayesian Skyline Plots. A new window opens and the Skyline Plots as images, which graphically shows the evolution and divergence of virus subtype in the selected geographical area and date of isolation.

What can go wrong:
The user’s firewall prevents the new windows from opening. The user should then unblock the webpage and allow for pop-ups from the website.

The system may crash. He should restart the system, and select a more specific search.

There may not be any cases for the specific query the user selected. The program will display a dialogue box that says: “No data available for selected query.”

**Other activities:**
All the results can be exported. The search can be saved.

**System State on Completion:**
New windows are opened with results from analysis showing. The user has yet to save or exported any of the results. When the user saves the results, he will get an e-mail confirmation.

---

**Scenario 3**

**Initial Assumption:**
The user has already filled out the following fields: virus: influenza, subtype: H1N, host: all, geographical location: United States and date of isolation: from today to ten years back. He conducted a multiple sequence alignment, Bayesian Skyline Plots, and a phylogeographic tree.

**Normal:**
The user decides that he needs to save the sequence alignment to conduct analysis in different models. In the sequence alignment page, he scrolls to the bottom of the page and clicks export result. A save dialogue box appears that allows the user to specify the name and it will save as a nexus format. Next, the user decides to save the Bayesian Skyline plots. He opens the window with the skyline plots and scroll down to save button. A save dialogue box appears that allows the user to specify what image format the user want to save the Bayesian Skyline plots. The user select JPEG format and he names the image accordingly and he clicks save. Thereafter, the user decides to save the phylogeographical tree. He opens the window with the phylogeographical tree, and finds the button to save at the bottom of the page, and clicks on it. A save dialogue box appears that allows the user to specify what image format the user want to save the phylogeographical tree. The user select JPEG format and he names the image accordingly and he clicks save.

**What can go wrong:**
The particular format that the user might save the result may not be available. The user should refer to the help ZooPhy page to re-visit what formats the data can be saved in.

**Other activities:**
The user can go back to the sequence list and analyze other sequences or he can export the results.

**System State on Completion:**
All the data is saved, there are no more analysis that can be done.
System requirements specification

1. User interface
   1.1. Virus name and subtype
   1.2. Virus host
   1.3. Geographical location
   1.4. Time period

2. Database
   2.1. NCBI Gene, Nucleotide Taxonomy
   2.2. EpiFlu
   2.3. NCBI Pubmed
   2.4. Biodesign
   2.5. GenBank
   2.6. Gisbank

3. Search results
   3.1. Virus list in Excel form
      3.1.1. Multiple Sequence Alignment
      3.1.2. Bayesian Skyline Plot
      3.1.3. Phylogeographic Tree

4. Future version
   4.1. BLAST
   4.2. Phylogenetic Tree

System models

The following are graphical depictions on how the system interfaces. This will help aid communication between the customer and the developers to aid to better fulfill user requirements.
1. Virus name and subtype
   - H1N1
   - Search Virus

2. Virus host name
   - Human
   - Search Host

3. Geographical location
   - United States
   - Search Maps

4. Time period
   - From 10/7/2000
   - To 10/07/2010

Save this Search

Share Search: Facebook Twitter

Figure 3: User input page
Figure 4: Example of the Taxonomical Tree

Figure 5: Sample host taxonomy tree after selecting H1N1 at virus subtype
Figure 6: Sequence Search Results (GISAID 2010)

<table>
<thead>
<tr>
<th>Isolate ID</th>
<th>Subtype</th>
<th>Passage</th>
<th>PB2</th>
<th>PB1</th>
<th>PA</th>
<th>HA</th>
<th>NP</th>
<th>NA</th>
<th>MP</th>
<th>NS</th>
<th>HE</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPI_ISL_31448</td>
<td>H1N1</td>
<td>2289</td>
<td>2285</td>
<td>2172</td>
<td>1732</td>
<td>1523</td>
<td>1415</td>
<td>983</td>
<td>946</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>EPI_ISL_31450</td>
<td>H1N1</td>
<td>2289</td>
<td>2299</td>
<td>2168</td>
<td>1733</td>
<td>1523</td>
<td>1412</td>
<td>984</td>
<td>850</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>EPI_ISL_31452</td>
<td>H1N1</td>
<td>2293</td>
<td>2299</td>
<td>2168</td>
<td>1732</td>
<td>1523</td>
<td>1415</td>
<td>984</td>
<td>850</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>EPI_ISL_31453</td>
<td>H1N1</td>
<td>2303</td>
<td>2299</td>
<td>2168</td>
<td>1731</td>
<td>1523</td>
<td>1412</td>
<td>987</td>
<td>850</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>EPI_ISL_31454</td>
<td>H1N1</td>
<td>2280</td>
<td>2303</td>
<td>2175</td>
<td>1733</td>
<td>1514</td>
<td>1418</td>
<td>987</td>
<td>850</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>EPI_ISL_31455</td>
<td>H1N1</td>
<td>2204</td>
<td>2301</td>
<td>2175</td>
<td>1732</td>
<td>1512</td>
<td>1411</td>
<td>984</td>
<td>850</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>EPI_ISL_32634</td>
<td>H1N1</td>
<td>MDCK2</td>
<td>2292</td>
<td>2274</td>
<td>2175</td>
<td>1731</td>
<td>1510</td>
<td>1410</td>
<td>983</td>
<td>844</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>EPI_ISL_31580</td>
<td>H1N1</td>
<td>C1</td>
<td>--</td>
<td>2299</td>
<td>2151</td>
<td>1732</td>
<td>1519</td>
<td>1413</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>EPI_ISL_63820</td>
<td>H1N1</td>
<td>P0</td>
<td>2288</td>
<td>2301</td>
<td>2151</td>
<td>1732</td>
<td>1514</td>
<td>1414</td>
<td>983</td>
<td>850</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Figure 7: UML Diagram of the system (refer to index for bigger image)
System evolution

The main fundamental evolution that the system will be designed for will be the component for the natural language processing. This consists of the following. When a phylogeographical analysis is queried by a user, and the information provided by the databases is not enough, natural language processing techniques will be used to obtain geographical information missing from the databases. The natural language component of the system will connect to NCBI PubMed. This will allow the system to search the literature automatically when a user queries the database.

Risk Analysis

<table>
<thead>
<tr>
<th>Risk</th>
<th>Probability</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>System server breaks down. This could lead to expenditure of resources and time in getting the server running back again. This will result in not being able to deliver the requirements at the desired deadline.</td>
<td>Low</td>
<td>Catastrophic</td>
</tr>
<tr>
<td>Due to licensing, legal and ethical issues the local database cannot be used. For example, human subject research or HIPAA compliance are not met.</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>The developer does not design the software that can be used for layman. The effect of background and posses a good amount of risk. For example, a researcher who is not skilled in using computer probably find it difficult to choose search from the hierarchical taxonomic tree representation.</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>If the developer is on a vacation and the server breaks down, then there is no one to fix the problem. To avoid risk, the developer can train all of the employers in assigned to the program in the maintaining of the database.</td>
<td>Low</td>
<td>Catastrophic</td>
</tr>
<tr>
<td>If the system not tested again and again regularly and to its maximum extent, risks are associated with its breakdown while in operation. e.g. If the system is tested for doing MSA for 100 sequences, but the user runs an analysis to do MSA on 200 sequences. In that case there can be chances of system breakdown. Would lead to loss of capital, time and resource.</td>
<td>Low</td>
<td>Serious</td>
</tr>
<tr>
<td>Risk associated with changes in system design and architecture. If the nucleotide sequences are pulled from the two databases GenBank and GISAID together. If the developer wishes to create a new database within ZooPhy which contains the non-overlapping sequences present in both databases together with its proper connectivity to the parent database. In this case, the</td>
<td>Moderate</td>
<td>Serious</td>
</tr>
</tbody>
</table>
system architecture needs to be designed again.

If during middle of the process of development one of the main developers leaves the project and a new person replaces him. It is going to be very difficult for the new person to work on already started work, as he doesn't know the project from scratch. Would need time to figure out the exact schema, documentation and user needs. Expenditure of time and resource.

<table>
<thead>
<tr>
<th>Moderate</th>
<th>Serious</th>
</tr>
</thead>
</table>

Software Limitation: Risk associated with the limitation of the working of the system should be kept in mind. For example, in the user interface where the user needs to select the geographical location, he chooses very distant country and tries to plot a phylogeographic tree of evolution of specific strain of virus in those two areas. It is most likely that the search will result in unsatisfactory results as there might be no cases where the evolution of specific that strain of virus took place at both these geographical location.

<table>
<thead>
<tr>
<th>Moderate</th>
<th>Tolerable</th>
</tr>
</thead>
</table>

**Risk Strategy**

Risk strategy was conducted for the risks that contained either a high probability or catastrophic effect.

<table>
<thead>
<tr>
<th>Risk</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizational Risks</td>
<td>Create a back up, even if it takes buying extra server space. This can be very useful in scenarios where the system breaks down. Having backup will eliminate the chances of losing data and query.</td>
</tr>
<tr>
<td>Staff Unavailability</td>
<td>Have another person who can fill in the shoes of principal developer. This will same time and the task can me met on time.</td>
</tr>
</tbody>
</table>
Project Plan
Below is a Gantt Chart with the project plan

<table>
<thead>
<tr>
<th>Task</th>
<th>Resource</th>
<th>End</th>
<th>Duration</th>
<th>% Complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elicitation of requirement Phase</td>
<td>Dr. Matthew Scotch</td>
<td>10/1/2010</td>
<td>14</td>
<td>100.0</td>
</tr>
<tr>
<td>User Requirement Document</td>
<td>GI/PO/AS</td>
<td>10/7/2010</td>
<td>11</td>
<td>100.0</td>
</tr>
<tr>
<td>Use Cases</td>
<td>GI/PO/AS</td>
<td>9/30/2010</td>
<td>6</td>
<td>100.0</td>
</tr>
<tr>
<td>Risk assessment</td>
<td>GI/PO/AS</td>
<td>10/7/2010</td>
<td>8</td>
<td>100.0</td>
</tr>
<tr>
<td>UML Diagrams</td>
<td>GI/PO/AS</td>
<td>10/5/2010</td>
<td>4</td>
<td>100.0</td>
</tr>
<tr>
<td>Initial Model drafter</td>
<td>GI/PO/AS</td>
<td>10/12/2010</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Hello World Assignment</td>
<td>GI/PO/AS: Individual</td>
<td>10/12/2010</td>
<td>10</td>
<td>50.0</td>
</tr>
<tr>
<td>Functional Specification</td>
<td>GI/PO/AS</td>
<td>11/2/2010</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>UML State and class diagram</td>
<td>GI/PO/AS: Indiv</td>
<td>10/19/2010</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>UML sequence diagram</td>
<td>GI/PO/AS: Indiv</td>
<td>10/26/2010</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Test Plan</td>
<td>GI/PO/AS</td>
<td>11/16/2010</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Initial test plan</td>
<td>GI/PO/AS</td>
<td>11/18/2010</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Milestone: Demo</td>
<td>GI/PO/As</td>
<td>11/23/2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final team Presentations</td>
<td>GI/PO/AS</td>
<td>12/14/2010</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Milestone: design presentations</td>
<td>GI/PO/AS</td>
<td>10/19/2010</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Test Plan

1. **Requirements reviews**
   Graphical representations and prototypes will be discussed with the customer to verify that the interface looks and its human computer interaction is as expected.

2. **Prototyping**
   Simple prototype will be created to test the query. To do this, we will work with Changjiang Mei. He is the database administrator for ZooPhy. We will conduct queries using the interface and the same query will be conducted by Changjian Mei. We will cross reference the results and validate that they are identical.

3. **Test-case generation**
   Once the interface is developed, it will be tested. All inputs will be tested by selecting different virus and analysing the drill down algorithms. This will also be done for the host. We will test the geographical location interface as well. After selecting the geographical location via Google Maps, the end query will be tested to verify that only the selected location is shown. Additional test must be done, to ensure that the user interface is compatible with all internet browsers, and different operating systems.
Appendix A: UML diagram
References
