Scenario Design for Evaluation of Visual Analytics Tools to Support Pandemic Preparedness and Response

Shawn Konecni, Georges Grinstein, Loura Costello, and Heather Byrne

University of Massachusetts Lowell

ABSTRACT

The VAST 2010 Challenge included a mini-challenge with a problem from the bioinformatics domain. The scenario created for the mini-challenge focused on a hypothetical pandemic outbreak involving a rapidly evolving fictitious virus. Outbreaks require rapid deployment of limited and temporary resources to mitigate potential damage to society and loss of life. Authorities and health care professionals utilize decision support and bioinformatics tools to ascertain the situation as it develops in order to properly allocate limited resources. The design of real world scenarios with their accompanying synthetic datasets enables users to develop and evaluate better visual analytics tools for this purpose. We have developed an introductory scenario whose solution encourages cross-domain participation to enhance creativity and establish a foundation for future challenges in this area.

KEYWORDS: visual analytics, pandemic, evaluation, synthetic datasets, scenario design

INDEX TERMS: H.5.2 [Information Interfaces & Presentations]: User Interfaces – Evaluation/methodology; J.3 [Life and Medical Sciences]: Biology and Genetics

1 INTRODUCTION

As part of the VAST 2010 Challenge [1], mini-challenge three included a problem from the bioinformatics domain. The focus of the mini-challenge was on a hypothetical pandemic outbreak involving a rapidly evolving fictitious virus. Synthetic genetic sequence and disease characteristic data was generated and provided to participants to help trace the origin of the disease and identify the most dangerous viral mutations. As intended in previous challenges which centered on largely intelligence analysis problems, the design of real world scenarios with their accompanying datasets enable users to develop and evaluate new visual analytics tools [2]. In the case of pandemic outbreaks, these tools can improve forecasting and decision making when preparing for and dealing with these catastrophes.

To adequately prepare for future pandemics, authorities must be able to manufacture effective vaccines ahead of time and provide rapid response services at the time of the crisis. A response requires the rapid deployment of a variety of temporary resources such as skilled health care staff, treatment centers, food supplies, additional vaccine doses, and in some cases anti-viral drugs [3]. Without adequate preparation, pandemics cause major disruptions to societal infrastructure and services, further compounding their disastrous effects [4]. To mitigate the damage, decision makers and health care professionals need to respond quickly and on short notice using limited resources. They need to be able to make sound strategic decisions using all available information, biological or otherwise, despite the fact that the information may be incomplete. To address these challenges, we need to continue to improve our decision support and bioinformatics tools.

skonecni@cs.uml.edu; grinstein@cs.uml.edu; loura_costello@student.uml.edu; hbyrne@cs.uml.edu

2 SCENARIO DESIGN

The scenario we designed centered on a patient admitted to a hospital with an unidentified illness. That patient later died after developing symptoms consistent with a pandemic outbreak in progress. As expected, an autopsy revealed the presence of the virus in the patient's bloodstream. Investigators and public health professionals need to understand the evolution of the current outbreak not only to provide intelligence for other parts of the challenge (such as the origin of the virus), but to understand the characteristics of the disease in order to mitigate its impact.

Two of the four challenge questions involve characterizing the distance between genetic sequences to reconstruct likely evolutionary paths and determine the origin. The second two questions require an approach to understand how the disease characteristics relate to new viral strains. In particular, the questions ask what changes, if any, in the genetic sequence make the symptoms more severe and make the virus more dangerous overall. This would help identify which strains need more attention.

The VAST Challenge caters to professionals from a wide variety of domains. As such, to make this challenge accessible to the community at large, compromises had to be made in scenario design. This approach was implemented in order to encourage cross-domain participation, thereby enhancing the creativity necessary to develop the tools and methodology for solving these kinds of problems. For example, the scenario focuses on genetic sequence information for only a single viral gene. This specific viral gene codes for a surface protein on the viral particle which can be assumed to affect the host through interactions with other protein molecules. However, without additional information, relationships between genes within the viral genome cannot be analyzed. Future iterations of the challenge will address such compromises and will introduce more complicated, better modeled and realistic scenarios.

3 DATASET GENERATION

Three datasets were provided to answer the challenge questions (figure 1). The first one contained genetic sequence information collected from viable mutants during the ongoing outbreak. The second one contained genetic sequences for viral strains native to a particular country or region. The third dataset contained supplemental information for a number of factors related to virulence and drug resistance. These datasets were largely created using Perl scripts with some manual adjustments to fit the scenario and tie in to other parts of the challenge.

Dataset One: Current Outbreak Sequences

(58 sequences)

>118
ATGTCACCGCCCTGCGCAGTTCATAGGGCCTCTCTTCGCCGGAACACGGGTCCTTTCTGGATGGTGAGGGTTGTGG
GAAAGACTTGTAGCCATAACGCATATCCAGATTTCTGGTACGCGCTCACGTATTCCGAGCGCGTGAAGTT ... TAG

Dataset Two: Native Sequences

(10 sequences)

>Central_Africa
ATGTCACCGCCCTGCGCAGTTCATAGGGCCTCTCTTCGCCGGAACACGGGTCCTTTCTGGATGGTGAGGGTTGTGG
GAAAGACTTGTAGCCATAACGCATACCCAGATTTCTGGTACGCGCTCACGTATTCCGAGCGCGTGAAGTT ... TAG

Dataset Three: Disease Characteristics

(58 records)

Sequence_ID	Symptoms	Mortality	Complications	Drug_Resistance	At_Risk_Vulnerability
2	Mild	Medium	Minor	Intermediate	Medium
15	Mild	Low	Minor	Resistant	Low
118	Severe	High	Minor	Resistant	High

Figure 1. Generated dataset format

From a randomly generated starting sequence, additional sequences were created via point mutations or single base changes. The rapid mutation rate of this virus is attributed to its genome type, which is considerably faster than other genome types due to lack of proofreading mechanisms during replication [5]. By leaving out insertion or deletion mutations, we kept the sequences at a fixed length and prevented the challenge problem from focusing on sequence alignment methodology, although that would have been more realistic. No silent mutations were used, so all mutations resulted in codons that code for a different amino acid. Additionally, only basic structural characteristics of the sequences were included in the problem. The positions of the point mutations were chosen at random, avoiding computations dealing with weighted genetic distance. So in this case genetic distance is simply the minimum number of substitutions required to convert one sequence to another.

For the first question, a number of mutations were repetitively introduced to the starting sequence to build a likely evolutionary tree for a set of native viral sequences. One of the nodes was chosen as the origin of the current outbreak, which in this scenario is a country in Africa. From this sequence, an outbreak strain was created within a reasonable genetic distance so that various tools (e.g. phylogenetic trees, diff utilities) would yield the same answer, thereby tracing the disease to the particular country of origin. From the starting outbreak strain, additional strains were created using one mutation per generation. In this case, a single mutation is defined as a one- or two-base substitution that differentiates one sequence from another.

For the second question, participants had to identify which patients admitted to a hospital likely contracted the virus from a person of interest. This type of information would support intelligence analysis in other parts of the challenge. The assumption was made that genetic sequence information is available for the patients in question and that a single viable sequence can be identified for each infected patient with no co-infection from multiple strains. The answer was determined by comparing the genetic distances between selected outbreak strains.

The third dataset contained five disease characteristics including symptoms, mortality, complications, drug resistance, and at-risk vulnerability. Symptoms referred to what a patient experiences when contracting the illness (e.g. pain, sore throat, vomiting, and tremors) and were categorized as mild, moderate, or severe. Mortality referred to the likelihood of death as a result of the disease and was categorized as low, medium, or high. Complications referred to other unfavorable conditions as a result of illness (e.g. deafness, spontaneous abortion) and were categorized as minor or major. Drug resistance referred to the viral mutant vulnerability to available anti-viral drugs and was categorized as resistant or susceptible. Finally, at-risk vulnerability referred to a disproportional effect of the illness on certain risk groups (e.g. children, elderly) and was categorized as low, medium, or high. The characteristics were generated randomly at first. Then, a handful of key mutations were chosen and the characteristics for all strains containing the mutation were adjusted to exhibit an increasing or decreasing trend. Some of the changes were subtle, making both visualization and analysis important in determining the answer.

The third question asked participants to identify the top mutations that lead to an increase in symptom severity. Our answer was determined by calculating Pearson's correlation of the mutation change to the increase in severity for all outbreak sequences. Thus the top mutations would show a net symptom severity increase for all sequences containing the mutation. To answer the question, participants needed to use a combination of statistical techniques and visual inspection similar to the way a molecular biologist might evaluate genotype-phenotype correlations in an evolutionary tree [6]. However, a purely mathematical solution may leave out some alternative solutions which might come up when using more complicated biologically-based algorithms to reconstruct evolutionary relationships. These

inconsistencies will need to be investigated further in the next challenge iteration.

The fourth question was similar to the previous one, but required participants to identify the top mutations leading to most dangerous viral strains. In a worst-case scenario, the most dangerous viral strains would cause severe symptoms, have a high mortality rate, cause major complications, exhibit resistance to anti-viral drugs, and target high-risk groups. The assumption used in the scenario was that all disease characteristics would be treated with equal weight. In reality, a pandemic response would treat these characteristics differently in order to mitigate the effect of a catastrophe and limit loss of life. To determine the answer to this question, an aggregated measure of all five disease characteristics was calculated and then the top mutations were chosen based on correlation to the aggregate measure. Again such refinements will be addressed in the next iteration.

4 Results of the Challenge

All teams answered the first two questions correctly. A number of teams answered the third question correctly, which dealt with the relationship between mutations and a single disease characteristic. Most teams struggled with the fourth question, requiring all five disease characteristics to be analyzed simultaneously. Awards and acknowledgements were given to submissions based on solution accuracy, overall design and analysis, innovative tool adaptation, process explanation, and novelty of the visualizations.

The intent of creating this introductory challenge was to engage the mainstream visual analytics community to create new and improved decision support tools for the bioinformatics community and specifically for effectively handling future pandemics. The resulting entries included a wide array of visualization tools and approaches. Some teams utilized existing tools and others built new ones. The clever adaptation of tools outside the bioinformatics domain demonstrated the advantages of designing flexible and robust visual analytics software. Some teams employed bioinformatics expertise in determining the solutions, but overall, teams adapted to the challenge with existing resources and skill sets. A sample of visualizations from the top submissions is shown in figures 2 and 3 [7, 8].

5 Conclusion and Future Work

Future iterations of the challenge will be designed to encourage more participation from members of the bioinformatics community while still trying to engage individuals from other domains. Although a more complex and realistic scenario would have been ideal, we believe this introductory approach was necessary as a first step towards a more organized interdisciplinary effort in the next round. Subsequent challenge questions and synthetic datasets will support several disparate approaches to reach the same conclusion without making substantial compromises in scenario design. These questions will also provide a gradient of difficulty that will allow various teams to submit more sophisticated solutions according to their backgrounds and skill level. Additionally, by publically releasing all entries along with benchmarks and solutions, future participants will be able to build upon previous experience. It is anticipated that the mix of cross-domain perspectives will result in the development of new and effective tools and methodologies as we saw this year. As future benchmark datasets are created, the complexity of the challenge will increase to more accurately model real-world pandemic outbreaks.

Acknowledgments

We would like to thank Catherine Plaisant from the University of Maryland, Jean Scholtz and Mark Whiting from Pacific Northwest National Laboratory, John Bodnar from SAIC, and Michael Graves from the University of Massachusetts Lowell for their assistance with challenge administration and scenario design. This work was supported in part by the National Science Foundation awards 0947343 and 0947358 and in part by the National Visualization and Analytics Center™ (NVAC™) located at the Pacific Northwest National Laboratory in Richland, WA. The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC05-76RL01830.

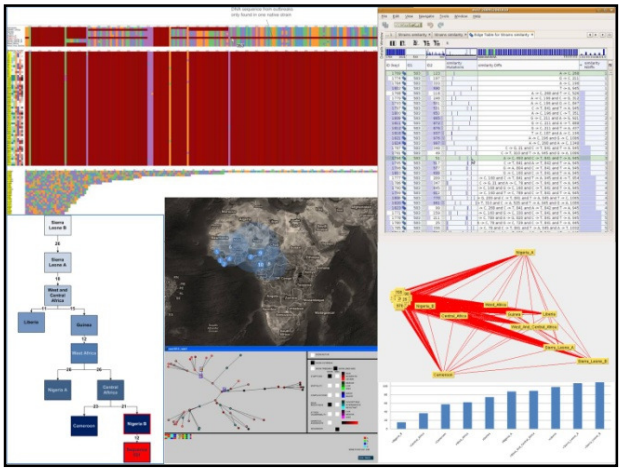

Figure 2. Sample of visualizations from question 1 and 2

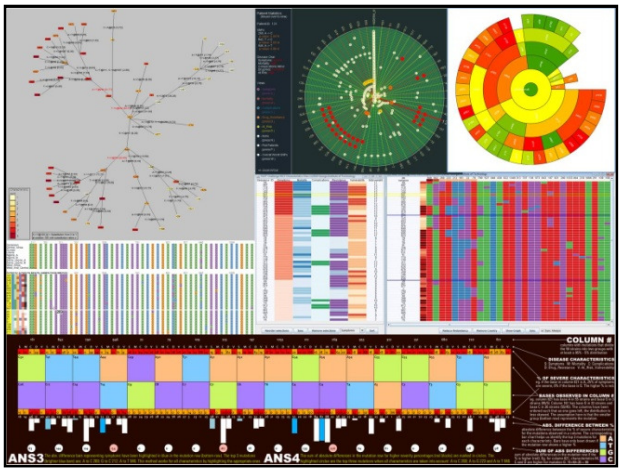

Figure 3. Sample of visualizations from question 3 and 4

References

[1] VAST Contest: http://www.cs.umd.edu/hcil/VASTchallenge2010/.

[2] Costello, L., Grinstein, G., Plaisant, C. and Scholtz, J., Advancing User-Centered Evaluation of Visual Analytic Environments through Contests, *Information Visualization* 8 (2009) 230–238.

[3] Osterholm, M., Preparing for the Next Pandemic, *New England Journal of Medicine* (2005) 352:1839-1842.

[4] Jenvald, J., Morin, M., Timpka, T., and Erikson, H., Simulation as Decision Support in Pandemic Influenza Preparedness and Response, *Proceedings of ISCRAM* (2007).

[5] Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S. and VandePol, S., Rapid Evolution of RNA Genomes, *Science* (1982) Vol. 215.

[6] Habib, F., Johnson, A., Bundschuh, R., and Janies, D., Large Scale Genotype-Phenotype Correlation Analysis Based on Phylogenetic Trees, *Bioinformatics* (2007) 23(7):785-788.

[7] Grinstein, G., Konecni, S., Plaisant, C., Scholtz, J., and Whiting, M., VAST 2010 Challenge: Arms Dealings and Pandemics, *Proceedings of IEEE Symposium on Visual Analytics Science and Technology* (2010).

[8] Visual Analytics Benchmark Repository: http://hcil.cs.umd.edu/localphp/hcil/vast/archive/.