**Genes4Vaccines: A computational model that utilizes comparative genetics to identify DNA & protein sequences for novel vaccines**

**Courtney Astore & Rebecca Elshansh**

**UCF Biotechnology**

**Jason Guo, David Durkee, Nicholas Kosan, Jayk Barker, Traci LaMoia, & Christopher Hodge**

---

**Introduction**

Pasteur was the first to develop a successful vaccine by inoculating chickens with attenuated forms of a bacteria. This resulted in the first generation of vaccines which were produced by isolating the pathogen, inactivating (killing) or attenuating it, and injecting into a person or animal. Second generation vaccines include inactivated vaccines and recombinant vaccines. Third generation vaccines use analysis of whole genomes to identify proteins and genetic sequences that can be used for vaccines, referred to as “reverse vaccinology.”

Genes4Vaccines falls under the category of “reverse vaccinology” and can be used for the following types of vaccines: rationally attenuated, subunit (including toxoid and conjugate vaccines), DNA, and recombinant vector vaccines. [6]

During the early stages of vaccine development, the risk of failure is at its highest. This is because much of early stage development is based off trial and error with different components for a potential vaccine [7]. In order to greatly reduce the costs of vaccine creation, it is necessary to reduce the amount of trial and error during vaccine discovery.

If the amount of trial and error was decreased, the speed of production would increase. This would not only be economically beneficial but it would also be beneficial to fighting pandemics. During a pandemic, it is essential to find a solution as quickly as possible in order to limit the amount of mortality or suffering in a population. During the 2009 H1N1 influenza pandemics, there was a delay in the vaccine supply.

The amount of trial and error can be decreased if a proper algorithm is necessary to find what data is necessary to make a useful vaccine and therefore limit the amount of trial and error in development.

---

**Methodology**

This pandemic showed that synthetic genomics techniques to produce influenza genome segments rapidly, accurately, and reliably are necessary in order to provide vaccines for production. Through the usage of genetic data, a vaccine can be developed and be produced by nearby laboratories. Thus, there would be no trial and error and vaccines can become more readily available because any lab can find the data and use it to produce vaccines. However, a proper algorithm is necessary to find what data is necessary to make a useful vaccine and therefore limit the amount of trial and error in development.

---

**Results & Discussion**

This model was developed to identify the significant gene starting and ending positions plays a role in the protein length. In order for Genes4Vaccines to successfully predict vaccine sequences, it is important to be able to predict the length of the gene. With this data, it is possible to increase the speed of vaccine production.

---

**Conclusion**

In its present state, Genes4Vaccines is not yet able to predict genetic sequences; however, within the coming two weeks, our preliminary algorithm will be completed and ready for analysis and optimization.

By constructing tree models, for the mass amounts and categories of data that have been collected, it has been possible to identify key properties to make comparisons with the target pathogen genome. In Figure 6, it can be seen that there is a strong relationship between the genetic sequence starting and ending positions and the success as a vaccine candidate. By filtering out markers with weak correlations and placing greater weight on markers with stronger correlations, the algorithms will have greater prediction accuracy.

Not only does the collected data contain sets that can be used for the prediction of genetic sequences, but there are also sets that can be used to determine the likelihood of success of the predicted sequences. With this, there was seen, not surprisingly, a strong relationship with length of the genetic sequence and success of the vaccine.

---

**Future Developments**

With the current algorithm, Genes4Vaccines focuses solely on predictions of antigen and virulence sequences. In the future, with identification of the proper databases, it will be possible to predict other properties. Each stage in antigen sequence, which will be the next immediate focus. The algorithms will also be adapted to predict tertiary and quaternary protein structures, opening up the realm of protein therapy.

In order to create a more viable product, it is important to have validation. To do this, both in vivo and test lab trials will be conducted. For in vivo validation, genomes of pathogens that have a vaccine, but that have been removed from the internal database, will be run through and the results will be compared to the actual genetic sequences used in the vaccine.

To further improve the statistical model, wet lab trials (in vivo and in vivo) will be done to further ensure the accuracy of the predictions and the ability to decrease the time spent on vaccine discovery.

---

**Acknowledgements**

- Aaron Smith, Ph.D, UCF Department of Mathematics
- Leslie Melon, Ph.D, Sanford Burnham Senior Director of Business Development
- UCF iCore Program
- National Science Foundation
- Female Thesis, Ph.D, UCF Biological Sciences Department

---

**References**

2. Center for Disease Control and Prevention. (2010). Genomes of pathogens that have a vaccine, but that have been removed from the internal database.