An evolution in the understanding of evolution

October 18 2019, by Chris Tyree

Engineering professor Kristen Naegle's new paper shows how to build models reconstructing evolutionary change much more accurately than ever before. Credit: University of Virginia

Remember domain, kingdom, phylum, class, order, family, genus,

species and Charles Darwin's "tree of life" metaphor we learned about in high school biology? That way of describing living-things lineages is just science's best guess about how genes have mutated and split over time to change things into what they are today.

It's not uncommon for living things to be reclassified into another genus as science gets better at identifying protein and gene changes; for example, there have been recent changes in taxonomy of different kinds of bacteria, plants and coral.

What if you could make a better model of evolutionary change that, while maybe not 100% accurate—considering complex organisms have been evolving for billions of years—could give you a clearer picture than ever before?

Kristen Naegle, associate professor of biomedical engineering and computer science at the University of Virginia's School of Engineering and resident faculty member of UVA's Center for Public Health Genomics, and her former Ph.D. student, Roman Sloutsky, now a post-doctoral researcher at the University of Massachusetts, Amherst, have done just that. Their work shows how to build models reconstructing evolutionary change much more accurately than ever before, which holds promise for breakthroughs in understanding how diseases work in the human body.

Their paper, "ASPEN, a methodology for reconstructing protein evolution with improved accuracy using ensemble models," was published Thursday in the journal *eLife*. ASPEN stands for "Accuracy through Subsampling of Protein EvolutioN." Their research highlights UVA's strengths in biomedical data sciences.

"Most models of protein evolution in use today are probably wrong," Naegle said. "We now have a way to poke at these models and ask how

we can use what is right about them to build better models. That's an important step."

To better understand the complex nature of their work in modeling evolutionary change, Naegle offers an analogy: "If I asked you to predict which route someone took between San Francisco and New York, that would be one model. But if I asked 1,000 people to give me a prediction of what route someone took, then the pieces of that route that are shared the most across all 1,000 people are most likely to be true. This is because most people might agree that a specific highway between two cities is the most efficient way to go, and so that section of highway would have a really strong weight, or probability.

"If I saw that no one agreed on anything across all those 1,000 routes, it would tell me I would have very little confidence in any one model being really accurate. Conversely, if everyone agreed on absolutely everything, or most pieces of the route, I would feel pretty confident there must be one best way to travel between those two points. I could come up with a new route that is not one that any of the 1,000 people gave me, but captures the most shared pieces of route between all 1,000 suggestions, and that model might be a whole lot closer to the true route than any individual model given to me.

"In the end, it still might not be wholly accurate—I can never know the real route unless I ask the person actually doing the traveling—but it's probably a lot better than any one of the route suggestions on their own.

"Evolution is like this, only it's like guessing a route through time instead of space."

Reconstructing evolutionary branches is tricky, especially when many species share a similar type of protein that might have evolved to perform somewhat different functions. Mathematically, the problem

quickly becomes very big, but discovering the implications of this protein evolution could lead to a better understanding of how our bodies deal with cancer and other diseases.

The solution to the problem came to Sloutsky while he was studying an important protein in cell signaling common to many different species. He wanted to know how the protein had evolved over time to have different functions in different species. The question was so big, he decided to sample just a few sequences to reconstruct the evolutionary divergence.

"The reconstructions didn't agree with each other," he said, despite 1,000 attempts. "That in itself wouldn't be a huge problem; I didn't expect them all to agree. But I expected one model to be repeated most of the time, or at least a lot of the time."

Surprised, he decided to see what all the disagreeing models had in common. "I knew I would have to come up with some way to combine information from all those models, because I couldn't just use the most common one," he said. "It was sort of an unexpected challenge that arose and led to this work."

Over the course of several months spent refining software and testing it on larger and larger reconstruction problems looking at proteins, Naegle and Sloutsky were able to create open-source software that can combine multiple models to very accurately reconstruct evolutionary changes.

"Everything our bodies do is done by proteins," Sloutsky said. "This is a powerful tool to understand how molecular biology works, how proteins work and when things go wrong, how they go wrong."

Naegle's and Sloutsky's raw data and code are included in the *eLife* publication so other researchers can use it for more precise modeling.

The journal *eLife,* focused on life and biomedical sciences, is unique among scientific journals. Peer reviewers assess the research and quality of the articles, and reviewers' questions and the authors' answers are included in the publication. The journal's philosophy is that knowledge should be open and accessible.

Researchers will be able to use Naegle's and Sloutsky's new tool, for example, to understand how highly similar proteins evolved and then design better drugs to target a protein more specifically. Naegle also imagines a physician trying to use medical imaging to discern the exact location and shape of a mass hidden deep inside a patient's body; this more accurate modeling tool could help the physician better understand the mass without cutting the patient open.

"George E.P. Box's much-quoted philosophy about models is relevant here: 'Essentially, all models are wrong, but some are useful,'" Naegle said. "We now have a quantifiable way to ask how good a model is, and by using the most useful parts across lots of models, we can build better models."

provided for information purposes only.