

Nathan Stouffer -- Attended Dr. Andreas Scherer's talk for CS Seminar

Summary

I attended a talk from Dr. Andreas Scherer, the CEO of Golden Helix. Golden Helix is a local bioinformatics company that provides services to medical researchers around the globe. They do a whole bunch of DNA processing and need to process data quickly. Dr. Scherer mentioned that there was something like 3.5 billion base pairs of DNA in a human genome (I'm real shaky on these bio terms so I could be spewing nonsense). The general premise of the problem they are trying to solve is how do we recognize problems in a gene string?

Dr. Scherer mentioned three 3 things that make this problem hard. The first is the scale of the input. As mentioned in the first paragraph, there is a massive amount of data that needs to be processed. Having an algorithm with a runtime of a high order polynomial is not an option, the processing algorithms need to be quick because there is so much data! The second is that the data is noisy. The method for reading a DNA sample is not entirely accurate, which makes it difficult to discern whether a DNA sample actually has a defective gene or is it just an error in the data collection process? The final problem is that some defective strings of genes are longer than the actual data samples. Dr. Scherer said that they are making great strides in methods that can look at an entire data string and tell if it is part of a longer defective string, which is pretty cool!

For much of the problem Dr. Scherer said that they use machine learning methods. This makes sense for two reasons. The first is that the data is noisy. With noisy data, it is difficult to build hard and fast rules for a typical algorithm to process effectively. There are just too many cases. Instead, we can create a model that finds the patterns in the data and trust that it is making a better decision than we would. The second reason this makes sense is that they are trying to solve a problem with partial input. This means, at best, a typical algorithm would have to have heuristics for accuracy and there is no way that we could prove the correctness of such an algorithm.

The service that Golden Helix provides is “automated clinical decision making.” This is the process of taking the DNA as input and returning a diagnosis of the patient as well as a recommended treatment. From what it sounds like, the company does not want to replace medical professionals, but instead aid their decision making process with a second opinion (of sorts).

Relationship to Algorithms

The talk did not explicitly talk about algorithms. However, the connection makes sense. DNA is necessarily discrete input and there is a whole bunch of it. This makes computers a great tool for processing such data. As mentioned in the summary section, a lot of the algorithms that Golden Helix uses rely on soft computing methods because of the variation in the data and the nature of the problems they attempt to solve. Machine learning methods are a form of algorithms, although not typically the type we study in this class so the relationship is there nonetheless.

Dr. Scherer did hint at some algorithms involved with string matching. The reason string matching is useful in this context is because there are some portions of human DNA that do not vary across all primates or maybe even all mammals. If there is a variation in those locations, then it is likely that there is a genetic defect that is present. Although, because of the noisy data issue, it seems that these algorithms would not be helpful on their own (enter machine learning!).

The Speaker

The speaker at the talk was Dr. Andreas Scherer. Dr. Scherer has a lot of experience running companies. From the bio that Sean sent out, he has been at the top of a few companies, gone through a bunch of school (getting his PhD), and even written some books. He seems like an ambitious guy with a bunch of good ideas.

Relevance to me

This talk was relevant to me because I will eventually make it out into the industry. It was cool to see what a company gets up to in Bozeman and get a high-level summary of their product.

Questions that were asked

Lucy Williams asked a question about the reference of a gene in regards to a certain population (I think, I should have written the word down). I'm not entirely sure what that means but it sounded interesting. Dr. Scherer said that her question pointed to an issue that is not entirely solved.