iGEM Heidelberg 2020: Prof. Dr. Fausto Giunchiglia

Professor for Computer Science
University of Trento, ECCAI fellow, member of the Academia Europaea

Three observations before answering the questions below. The first is that I am a computer scientist with some experience in working in the Health domain with very, very minor experience in computation biology. Because of this, you have to take what I write with a grain of salt. The second is that all the topics below will generate top quality publications. My answers below will therefore focus on whether they will have an impact (of course in my humble opinion). The third is that I'll distinguish the impact in the: (i) short term impact (which will happen in the next 1-2-3 years) and (ii) long term impact, namely something which will substantially change synthetic biology in the long term.

1. In iGEM (international genetical engineering machine), we are working with complex biological systems. Do you see any benefit in applying formal methods (e.g. model checking) to design robust biological systems?

I have a long term experience in formal methods and I have contributed to the development of one of the most commonly used open source model checking platforms, i.e. NuSMV. I consider it unlikely that these techniques and methodologies will have a short or long term impact in this area (they had and still have a short and long term impact in other areas). This is because any such methodology assumes a model of computation that I do not believe can be extended to model biological systems. Adding probabilities and similar things will not do it. We need a much deeper revision of the underlying computational model (where, of course, the computational model is how people model cause-effect).

2. On the software side, we are working with maximum-likelihood language models applied to protein and RNA sequences. In which way do biological sequences differ from natural language? In which way could we treat them similarly? How could modelling of protein sequences benefit from the wealth of knowledge in natural language processing?

Natural language has a sequential structure. This is its essential backbone. Then, on top of this, one can construct secondary connections. Not being an expert in biology, it does not seem to me that in this field there are similar structures. If I had to choose an approach I would focus on knowledge graphs and graph embeddings. Graphs are much more flexible and can be used to model any time of connection, e.g., spatial, temporal, cause-effect.

3. Your experience in applying machine learning to natural language data is very impressive. What are some pitfalls to training such machine learning models?

Machine learning applied to NLP is very powerful. The pitfalls are:

- In particular if you use DNN, it is hard to understand how the decision was made. The machine does not know whether she got it right or wrong and it is still an open issue how humans can recognize it (also given the huge number of states of the algorithm)
- You need to have lots of examples. Do you have them? How do you know they are not biased?
- The problem of how to use machine learning and DNNs in time is unsolved. One of the problems that people have now is how to deal with what is called *catastrophic forgetting.*
- How to make sense, in human terms, of what machine learning produces. A lot of work is now focusing on how to integrate with more knowledge based approaches and on what people call the human-in-the loop.

All in all, I guess it is fair to say that DNNs work very well in all those applications where the cost of a mistake is low, or where there is very little prior knowledge. Their extension to more mission critical applications is still an open issue. They work in general well with images and show more difficulties with symbolic knowledge (just for the reasons above).

4. When we are modelling proteins, in addition to their sequence we are commonly given structural data (i.e. their 3D structure and contacts between amino acids). Are there parallels in natural language processing for explicitly representing long-range dependencies in natural language?

Use knowledge graphs and graph embeddings. That to me is the way to go.

5. One aspect of language processing commonly applied to biological sequences are (context free grammars). Are there other concepts from natural language processing adjacent computer science you think could be beneficial to treating biological sequences?

Language can be seen as a set of sentences which represent in different way similar or identical meaning. The techniques used to disambiguate language are strongly influenced by the fact that they focus on language. Repeating myself, I think that graphs is the way to go. They can just be shaped to model exactly the phenomenon of interest.

6. Which aspects of your work do you think could be applied to biological sequence data?

I study the process of how entities adapt and evolve in time, also as function of the local state, thus changing radically. I also study how this “diversity” can be “absorbed” in a single representation when you need to put together different entities, or different representations of the same entity. I use knowledge graphs to represent diversity. Not being an expert in your field, and assuming that you also need to model change and cause-effect, I am pretty sure you have this kind of problems.

7. There is currently a lot of hype about machine learning in general. What is your opinion on this?

It was a major success. It has changed the way we do anything, not only computer science. Now it is just makes no sense to do without it. But this research is now in a plateau as it run in a set of foundational problems (some listed above). We need something radically new to enable the next big jump. But I guess the time is ripe for its applications in field like this one.