Protein Linguistics

moalquraishi.wordpress.com/2018/02/15/protein-linguistics

February 14, 2018

For over a decade now I have been working, essentially off the grid, on protein folding. I started thinking about the problem during my undergraduate years and actively working on it from the very beginning of grad school. For about four years, during the late 2000s, I pursued a radically different approach (to what was current then and now) based on ideas from Bayesian nonparametrics. Despite spending a significant fraction of my Ph.D. time on the problem, I made no publishable progress, and ultimately abandoned the approach. When deep learning began to make noise in the machine learning community around 2010, I started thinking about reformulating the core hypothesis underlying my Bayesian nonparametrics approach in a manner that can be cast as end-to-end differentiable, to utilize the emerging machinery of deep learning. Today I am finally ready to start talking about this long journey, beginning with a preprint that went live on bioRxiv yesterday.

End-to-end differentiable learning of protein structure

What ultimately doomed the Bayesian nonparametrics approach was computational cost—I spent millions of compute hours on the problem without making any headway in getting the models to converge. I thought (and still do) that that formulation captured something fundamental about protein folding, but the mathematical tools for sampling and variational inference were and are not yet mature enough to make it work. Deep learning presented the appealing possibility of casting protein folding as an optimization problem that can be optimized end-to-end using gradient descent. I knew very little about neural networks back then but started studying them seriously in 2012. During the period spanning Sept. 30th, 2014 to Feb. 18th, 2015, I settled on the basic formulation that I would eventually call Recurrent Geometric Networks or RGNs (I track things). It took another three years from conception to realization—deep learning frameworks were nowhere near as mature then as they are now, causing me to start over with a new codebase twice, and RGNs can be quite challenging to train. But, I think they are now finally ready to be released into the wild! For the technical details, go read the preprint. In this blogpost, I'd like to describe the thinking process that led up to them.

From when I first learned about protein folding, and the approaches taken to predict protein structure, I thought it may be possible to predict proteins without conformational sampling and energy minimization, the two pillars of protein structure prediction. The reasons for this have come to underlie what I call the linguistic hypothesis. The basic idea is as follows: there is evidence that today's proteins emerged out of an ancient peptidic soup, one that may have left its mark on the evolutionary record. I.e., the proteins we see today may in some sense be formed out of primordial peptides. As proteins grew in size and complexity, it would have been advantageous to reuse existing components, to build bigger proteins from existing protein parts. We already know this is true on the level of protein domains, in that larger proteins are often comprised from chaining together smaller globular domains. But the phenomenon of reuse may go further, where even smaller protein fragments (handful of residues to dozens) may reflect an underlying evolutionary pressure to reuse working parts, fragments that fold in tried-and-tested ways (from the perspective of evolution.) If this is the case, then the space of naturally occurring proteins may occupy a very special "manifold", one that exhibits a hierarchical organization spanning small fragments to entire domains. Other evolutionary pressures could further drive the reuse phenomenon. For example, once a protein-protein or protein-DNA interface is established, presumably through some sort of structural motif, reusing that motif would present an efficient way for the cell to rewire its cellular circuitry. The end result of all this would be the emergence of something resembling a linguistic structure, a grammar that defines the reusable parts and how these parts can be combined to form larger assemblies. Given that this is biology, it's unlikely to be rigid or minimal. It would be messy and hacky, with many exceptions and *ad hoc* evolutionary optimizations. But the manifold would be there, potentially discoverable and *learnable*.

What to me is most exciting about this idea is the emergence of a layer of biological phenomenon that is rooted in physics (obviously), but that can be described independently of it. I.e. the primitives of this phenomenon would be sufficiently abstracted away from the underlying physics that we don't need physics to model it. We can just operate on the level of protein fragments and motifs, building a probabilistic grammar that describes how these parts combine and interact, without ever resorting to a brute force physics-based simulation. To some this may seem unprincipled. To me, this possibility is *more* exciting than being able to brute-force simulate protein folding, as such simulations would ultimately only be exercises in doing physics really well and at scale. But the possibility of the existence of a description of protein structure space that can be formulated without resorting to the underlying physics, is the possibility of a theoretical science of proteins, independent of (but ultimately rooted in) physical theory.

My original Bayesian nonparametrics approach very explicitly codified this intuition. That is the advantage (to me) of Bayesian nonparametrics. They are a great way to encode one's prior about the *generative* process underlying the phenomenon of interest. In some ways, designing neural network architectures goes in the exact opposite direction. Instead of capturing our prior about the underlying generative process, deep learning works by capturing our prior about the inverse learning process. I.e. the architecture encodes how the phenomenon can best be learned from data. Consider computer vision: images are not actually generated by convolutions, by having copies of cats repeated across our visual field! But, given a natural image, we know that a good prior for learning patterns in images is the imposition of translational invariance. It's a statement about what makes a good learning process, about regularities in the phenomenon of interest that can be exploited by our learning process.

Recurrent geometric networks try to do this for the linguistic hypothesis I articulated above. There's nothing in them about ancient peptides or reuse. But they are structured in a way to discover patterns in protein sequences, and to discover hierarchies of such patterns. Perhaps more importantly, they are set up so that the link between sequence and structure is direct and immediate. The signal for whether the learned representation is useful for protein structure prediction comes directly from the predicted structures themselves, because they are explicitly compared to real structures and the deviation between the two is backpropagated to the learned weights of the representation. The hard or what one might call the clever part of RGNs is making this happen, the coupling between the representation and the final output. Ironically, I think the idea itself is very simple and straightforward. I had it over three years ago and it always struck me as very obvious. The real hard part in many ways has been getting it to work.

So are RGNs a panacea? Not at all. This is very much a 1.0 release. They are raw and unpolished. Training them can be quite challenging, like I already mentioned. They do comparatively well on novel protein topologies, but that's because everyone else does so poorly. They do silly things like predict pretty awful secondary structure, and their predictions can have steric clashes and the like. The specific preprint I just posted has some lame aspects—for example it uses the CASP11 structures as a testbed, instead of the more recent CASP12, for no other reason than the fact that when I first started training them, CASP11 was still current!

But, and this is the key point, I don't think any of this really matters at this early stage, or distracts from what's most exciting about them. RGNs can predict protein structures, at a very competitive level, without sampling! Without energy minimization! Without templates! And without the key driver behind the recent successes of protein folding, co-evolutionary data. I was intentionally somewhat puritanical in this paper, in the sense that I didn't add any bells and whistles such as physical priors, templates, or co-evolutionary information, because I wanted to communicate the key finding that even without all these things, i.e. while being orthogonal to what *currently* makes protein structure prediction work, RGNs can do pretty well. This means, I am rather certain, that with enough engineering, perhaps Google-scale engineering, it would be possible to make RGNs work *really* well, maybe shockingly so. To be sure, that's speculation at this point. But I think it's clear that they're a very different way to model protein structure. RGNs reason about proteins in a way distinct from the kind of computations done by molecular dynamics, or fragment assembly, or certainly the sort of optimizations done to extract contact maps from co-evolutionary data. And that to me is fundamentally very exciting.

I hope you take a look and find it as exciting as I do. If I'm successful, this is the beginning of something new, rather than the end of anything. Protein folding has yet to be solved, but we're living in the most exciting era of this foundational problem; one that may have us see its resolution.