A DEEP DIVE INTO MODELING AND MECHANISMS

Interview with Dr. John Moult

BY BRYAN HSU, TIMOTHY JANG, AND ANANYA KRISHNAPURA

John Moult, PhD, is a professor in the Department of Cell Biology and Molecular Genetics at the University of Maryland. As the principal investigator of the Moult Group, based at the Institute for Bioscience and Biotechnology Research, Dr. Moult focuses on the computational modeling of biological systems. Dr. Moult is also the co-founder and president of the Critical Assessment of Methods of Protein Structure Prediction (CASP) challenge, a long-running protein modeling competition that aims to advance methods of predicting protein structure from sequence. In this interview, we discuss the growth and development of the field of protein modeling and prediction, as well as Dr. Moult's work towards creating a framework for modeling biological mechanisms.

BSJ: What initially fueled your interests in protein modeling and the field of computational biology as a whole?

JM: In some sense, it was somewhat of an accident. The initial experimental work on protein folding was done by a man named Christian Anfinsen in the early 1960s. I became a graduate student in 1965. At one point, my supervisor asked me, "Why don't you go and solve this protein folding problem in your spare time?" I was not able to do anything about it at that time, but I did get intrigued by the problem, and gradually this interest had more and more of an impact on the research I did later.

BSJ: What are some of the challenges associated with determining the 3-D structure of proteins?

JM: From a computational point of view, determining the 3-D structures of proteins comes with three main difficulties. First is the size of the search space, which describes the set of possible orientations of a protein. An unfolded polypeptide is very complex, and you can think about the number of ways it can be arranged in 3-D space in many different ways. One way I like to think about it is how DeepMind [subsidiary of Alphabet, Inc., whose research aims to construct AI systems] puts it. They compare the complexity of an unfolded polypeptide chain with that of the game of Go, where you have 19 by 19 positions. There are three states for each position: blank, black, or white. Therefore, there are 3361 total possibilities. You can think of a polypeptide chain with the same number of amino acids (361) as having roughly the same number of possible conformations if you approximate that there are only three conformations per amino acid; however, even this is an underestimate.

The second difficulty comes from the folding of the protein. As a protein begins to fold and become compact, an increasing number of limitations on its movement arise since polypeptide chain bits cannot move through each other. The third difficulty comes from the fact that the free energy difference between the unfolded and folded states is small compared to the individual interactions between atomic groups within a protein. These three difficulties combine to make it computationally challenging to determine protein 3-D structure.

BSJ: You co-founded the Critical Assessment for Structure Prediction (CASP) challenge in 1994. At the time, what was your main intention in creating this challenge?

JM: Our purpose was to try to accelerate progress in solving this one problem: computing 3-D structure from sequence. The issue can be summarized as follows: when doing work in a virtual world, one can create a digital twin of a protein, but in doing so, this virtual world gives the individual too much leeway as to how they can arrange the protein. You end up losing the normal rigor of experimental science. The idea of CASP was to come up with a simple way of putting the rigor back into the process.

There has been tremendous progress toward this goal since we started the challenge. We can now employ several techniques in protein modeling. For example, modeling using homology to other proteins has had a lot of steady progress and has become a very useful

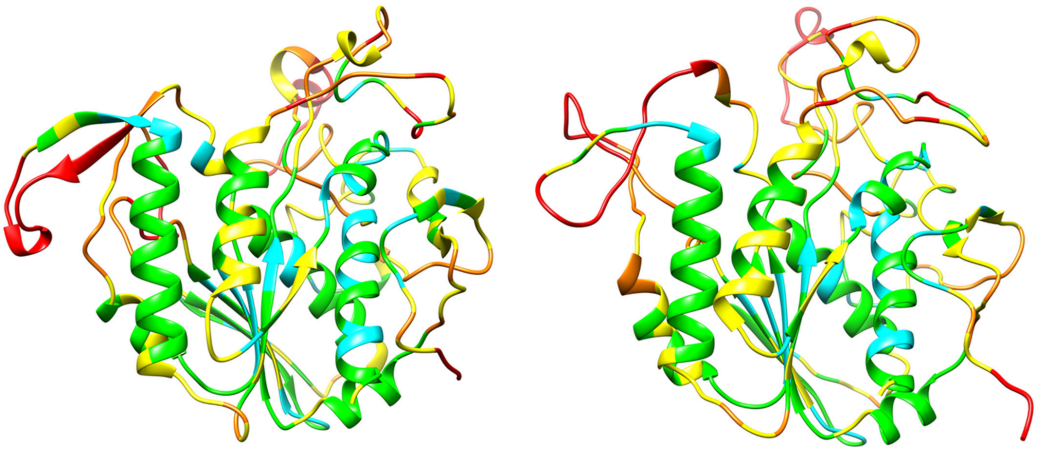


Figure 1: The left panel illustrates the crystal structure of the 354-residue domain ESKIMO 1 (6CCI) compared to its most accurate CASP13 prediction model on the right. *Colors in the image represent the accuracy of the model, showing high accuracy in the core with green and blue and lower accuracy in the edge loop regions with orange and red. Image copyright © 2019 by Wiley Periodicals, Inc; reprinted under fair use.*

technique. The more fundamental problem of, "Can you calculate the structure without using much homology information?" has proven much tougher. We have seen several incremental improvements over the years, but it is in the past six years that certain methods have really taken off in order to address this problem.

BSJ: What is deep learning, and how has it altered the way we predict biological structures?

JM: Deep learning has emerged from the application of neural networks or other machine learning methods to predict experimental outcomes from data. While practical applications of modeling protein structure using machine learning began in the 1990s, these methods were highly restricted because of technical difficulties with the algorithms. Around 2010, there were a series of breakthroughs in addressing this problem. One allowed for the ability to construct much bigger networks. The word "deep" in deep learning actually refers to the number of layers you have in a network. Whereas having more than three layers was previously considered "deep," there are now networks with hundreds of layers. With deep learning, rather than having to pre-process the data, you can give the network raw data and it will sort out what is important.

BSJ: Recently, there has been a lot of excitement about one CASP14 project in particular that utilizes deep learning: AlphaFold 2. How does it work differently from prior approaches?

JM: There are some significant differences from other approaches. Earlier, I mentioned that there have been several major improvements in the field of protein modeling over the past six years. The first thing that happened was that traditional statistical methods became successful at predicting which amino acids are in contact in the protein's three-dimensional structure. These predictions now provide some restraints to the possible 3-D structures you can predict for the protein. Then, about four or five years ago, people recognized that you could represent the set of contacts between the amino acids in a folded protein as an image. To do this, you make an amino acid sequence by an amino acid sequence array with n^2 pixels, where n is the number of amino acids in the sequence. You fill in a pixel if there is contact between the amino acids. Otherwise, pixels remain blank. The result is a two-dimensional image. This is where deep learning comes in, and it has been very successful with image recognition. In a previous CASP round about two years ago (CASP13), a number of groups began applying these ideas—treating the contact maps as images and training convolutional neural networks to successfully predict the folds of most proteins. This was a very exciting advancement since the general topology of most proteins could be correctly modeled. However, atomic details were still elusive.

In that CASP, DeepMind was the most successful, which was very impressive since they came in from outside the field. They built on the ideas that others had already developed within the community. In the following two years, they realized that they were stuck; with the way they were approaching the problem, they were not going to get to atomic-level accuracy in their models. So, they abandoned most of the previous technology and, as they say, explored at least a dozen other methods through the prototype stage. As a result, a few critical algorithmic changes were introduced. The first was getting rid of using a convolutional framework, which is not ideally suited to these types of images. Instead, they decided to use the currently emerging technique of attention learning. The second change they made was to the final stage of the network. Rather than outputting a set of predicted contacts, the network's final stage now produces three-dimensional coordinates through the use of newly emerging technology. They also built some protein properties directly into the network structure. These changes resulted in advancing us from getting the fold right two years ago to now achieving atomic accuracy.

BSJ: What do you aim to address as the future targets of CASP_Commons and CASP15?

JM: CASP_Commons has been an attempt to, rather than test how well methods work against an experiment, actually

> "These changes resulted in advancing us from getting the fold right two years ago to now achieving atomic accuracy."

use the CASP community to address issues of significance. It has been quite successful recently in producing models of some of the most difficult-to-get SARS-CoV2 structures. Now that we have experimental data, we can confirm that some of these predictions have turned out to be very good. Of course, we are going to try to keep pursuing this line of research, but the main goal of CASP is to advance methods for protein modeling. Our real next challenge is predicting the structures of protein complexes. The folding problem was defined a long time ago when we thought about single proteins, but we now know that biology is really all about protein-protein interactions, and we want to be able to predict these. There are methods which already display some progress on this problem, but they do not quite nail down a solution. The expectation going forward is that the same sort of deep learning I mentioned previously, along with some tweaks, will be successful at solving this problem. Seeing whether this approach will be successful or not is the real source of excitement for the next CASP.

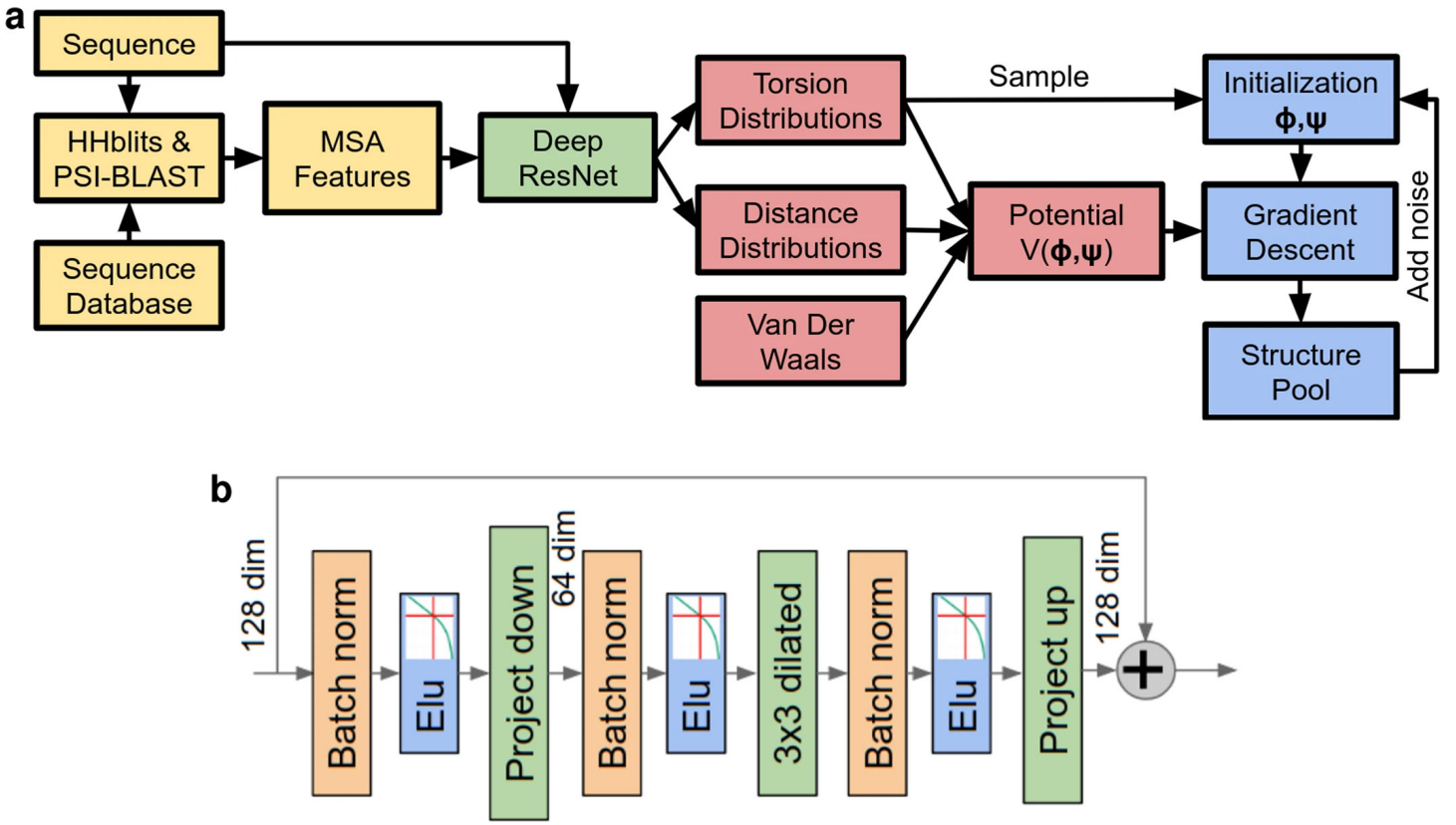
BSJ: Much of your research is currently focused on the application of computational methods to model not just proteins, but biological systems. What drew you to computational biology?

JM: I got a taste for doing things computationally with my work on the protein structure problem, and from there, I began to see that the broader field of computational biology would be central to the future of biology. I think it is going to be like theory in physics in that it is going to drive the experimental process. In terms of my specific pathway into the field, my lab and I were initially interested in the question of how genetic variants affect disease.

BSJ: You have helped create MecCog, a framework for representing biological mechanisms, especially those connecting genetic variants with disease outcomes. What prompted MecCog's creation?

JM: The reasons are similar to those that led to CASP's creation. The institutions and the procedures individuals have created in order to do science are really amazing, but they are not perfect; they have a sort of inertia. As the science we do changes, the way that we go about it does not change fast enough. One field of study this particularly applies to is the study of complicated biological systems. Individuals are not able to measure exactly what they want to measure and do not have a way to efficiently organize existing data. For example, my lab and I are interested in Alzheimer's disease, where the focus is the human brain, something as inaccessible as you can get in terms of measurement. Of course, you can do autopsies, scans, and so on, but you cannot really make molecular measurements in living human brains. Instead, you have to make measurements in mice or in cells and extrapolate from those measurements the mechanism behind the disease. In principle, that is not difficult, but in practice, it gets very, very messy.

Figure 2: Schematic of a previous generation of DeepMind's protein structure predictive system, AlphaFold. *The structure-prediction neural network is shown in green. Specifics for the most recent version of the network have not yet been released publicly. Image copyright © 2020 by the authors; reprinted under fair use.*



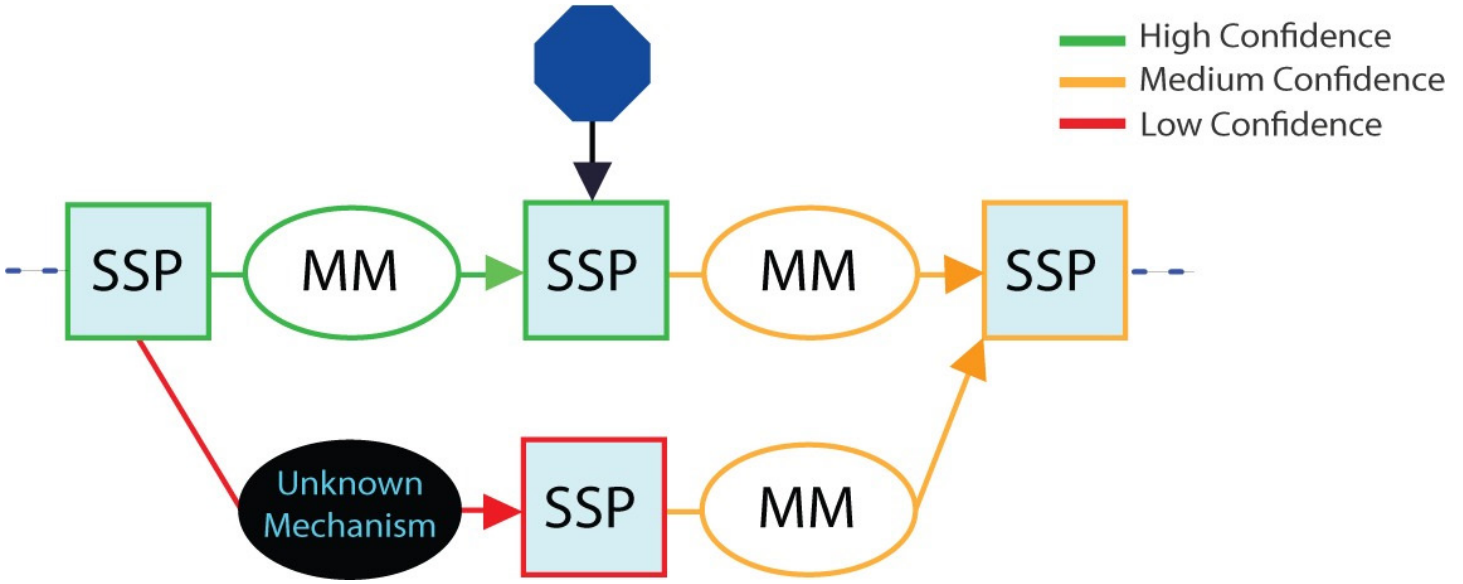


Figure 3: General layout of the MecCog disease mechanism schema. *The schema shows decreasing levels of certainty in evidence with the colors green, yellow, and red respectively. Black circles represent unknown mechanisms, and blue octagons indicate possible sites for therapeutic intervention. Mechanism modules are abbreviated as MM and substate perturbation as SSP. Image licensed under CC BY 4.0.*

Now, in Alzheimer's, there is one gene variant called APOE4 that predisposes you to developing the disease. If you have two copies of this variant, you are around 30 times as likely to develop Alzheimer's compared to if you had the normal variant. Obviously, there has been a lot of interest in this one variant, and there are around 10,000 published papers on the protein APOE. However, if you wanted to treat patients with this disease variant, you would not know whether it would benefit the patient to have more or less activity of this protein. We have 10,000 papers on the protein, and yet we still cannot answer the most basic qualitative question. Why is that? It is because of the remoteness of the experiments and the inability to organize the information you do have. MecCog is an attempt to come up with a framework that does exactly that; it solves this problem of how we can systematically think about mechanisms in a way that helps us sort out what we know, what we do not know, and what experiments to do.

BSJ: In the language of MecCog, what are "entities" and "activities," and what role do they play within a mechanism?

"The institutions and the procedures individuals have created in order to do science are really amazing, but they are not perfect; they have a sort of inertia."

JM: We can think of mechanisms, particularly disease mechanisms related to genetic variation, as a series of steps. You start with the DNA, which is an "entity," and you perturb it; this could be a base change in DNA. In the language of MecCog, this change is called a "substate perturbation," where the state is the state of the DNA. Then, a mechanism module or "activity" links this change at the DNA level to the protein level. For this example, transcription and translation would link this base change in DNA to changes in the amino acids of a protein.

Essentially, an entity that is perturbed, in this case the DNA, is linked by some activity to a change in another entity, such as a protein. We can thus think of a mechanism as a string of perturbed entities linked by a string of perturbed or normal activities. Of course, causal networks are fairly well established, but the difference with MecCog is that you can label the edges in a causal network with these mechanism modules or activities.

BSJ: Could you give us an example of how one might analyze the interaction between genetic variation and disease phenotype through MecCog's framework?

JM: My previous example on Alzheimer's and the APOE4 variant really illustrates the benefits of using MecCog. Normally, if you think about how a base change affects disease phenotype, there might be one or at most five different mechanisms linking the two. However, by my count, 22 different mechanisms have been proposed and supported by data for how APOE's base change affects the risk of Alzheimer's.

Let us take one of these mechanisms as an example. Within this mechanism, there are two different inputs: the APOE4 base change and environmental perturbation. This environmental perturbation refers to stress on neurons, which happens under injury conditions

or with age. Stressed neurons produce more of this APOE4 protein, which is less thermodynamically stable than the normal protein. The less-stable protein, which is an entity, is more susceptible to proteolytic cleavage into two pieces. In turn, that allows the mechanism module of cleavage to go faster than it does with the normal protein. Thus, the next altered entity or substate perturbation in this mechanism is this state of having more cleaved protein than you would get with a normal protein. The cleaved protein goes on to bind to a protein called tau, one of the major players in Alzheimer's. Tau is normally associated with microtubules, but its interaction with the cleaved APOE protein appears to detach it from microtubules. Ultimately, the increased aggregation of tau is one of the drivers of neural deterioration.

BSJ: As of now, data for mechanism schemas must be manually inputted by researchers. How do you see this process changing over time?

JM: When you do something like read a paper and say, "Okay, the relevant event for this mechanism step is the cleavage of this protein," your mind has extracted information from the paper and formalized it into a MecCog-type arrangement. Currently, there are no methods for automatically replicating that. Of course, there is a huge amount of AI work going on in interpreting language. However, if you actually look at what has been achieved in the biological sciences, it is pretty disappointing. Current methods cannot even succeed in reliably identifying which proteins are referred to in a given text. So we are a long way from automating this process and directly mining information from papers, but it is really intriguing to think about how you might do this. What is happening now is this application of things called transformers to language, where you essentially transform the relationship between the words in order to make them nearer to the concepts. I am not sure, but maybe something like that is going to have an impact here. One thing this all emphasizes is the strange way in which we use language. How we use language in science is a very flexible and powerful thing, but its very flexibility makes it hard to deal with computationally.

BSJ: Finally, what kinds of developments do you foresee MecCog leading to within the biological sciences, such as the medical field?

JM: With 10,000 papers, I believe there is no way individuals or research groups can make sense of all the information out there or accurately model 22 schemas for each proposed mechanism. MecCog is designed as a crowdsourcing tool. For key diseases like Alzheimer's, we are hoping to build a repertoire of all of these mechanisms to see what we know and what we do not know. Once we have these things laid out, we can then think of potential therapeutic strategies.

What serendipitously came out of this was that if you have multiple inputs into a disease, you can draw an intersecting graph of all the different schemas. This results in a sparse neural network connecting genetic inputs to a disease output. If you have enough data, you can then train this model to predict disease state from an input. Additionally, if you have the topology of the model right, you will generate the right functions at each node in that sparse neural network. For example, we constructed a network to model part of Crohn's disease. Six DNA variants in the input for this network contribute to an unfolded protein response, represented by one of the internal nodes. The network correctly learns the complex relationship between these inputs and elicits a response depicting a sigmoidal set-in. We did not tell the network that this is the sort of set-in it should give, but it learnt that this is what happens at that node. And now, with this network, you can ask, "If I were to drug the patient at one of these nodes, how would the network respond? Would administering the drug decrease the probability of disease?" Because we have so much data, in the future, I envision that we could have something loosely analogous to deep neural networks representing biological systems. To me, that is the really exciting way forward.

"How we use language in science is a very flexible and powerful thing, but its very flexibility makes it hard to deal with computationally."

This interview, which consists of one conversation, has been edited for brevity and clarity.

IMAGE REFERENCES

1. *Headshot:* [Photograph of John Moult]. Institute for Bioscience & Biotechnology Research. https://www.ibbr.umd.edu/taxonomy/term/445
2. *Figure 1:* Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins*, 87(12), 1011–1020. https://doi.org/10.1002/prot.25823
3. *Figure 2:* Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7
4. *Figure 3:* Darden, L., Kundu, K., Pal, L. R., & Moult, J. (2018). Harnessing formal concepts of biological mechanism to analyze human disease. *PLOS Computational Biology*, 14(12), Article e1006540. https://doi.org/10.1371/journal.pcbi.1006540