Philippe Youkharibache, Ph.D. and Jiyao Wang, Ph.D.

Behind the Mask

October 1, 2021

Barr: Good afternoon. Today is October 1, 2021. My name is Gabrielle Barr. I'm the Archivist at the Office of NIH History and Stetten Museum. Today I have the pleasure of speaking with Dr. Philippe Youkharibache, who is a Staff Scientist in the Cancer Data Science Laboratory at the National Cancer Institute (NCI), and Dr. Jiyao Wang, who is a Staff Scientist at the National Center for Biotechnology Information (NCBI), which is part of the National Library of Medicine (NLM). They are going to be speaking about a resource they have been creating, iCn3D [I See in 3D], and how it relates to COVID-19. Thank you for being with me.

Wang: Thank you.

Youkharbache: Thank you.

Barr: When did you all begin developing the iCn3D resource—which is a really great title, by the way—and can you describe the context behind your creation?

Wang: [We started] around 2015. We previously used Cn3D [see in 3-D] and then we found that the web version of 3D viewers such as Jmol watcher actually works pretty well. We tried to use the available options, like 3Dmol and Jmol, but we found that that software, at that point, didn't satisfy our requirements. We decided maybe we could try our own. Because it's a JavaScript viewer, it's open source. That's how we get started.

Barr: Can you speak a little bit about some of the hackathons that helped shaped the resource?

Youkharbache: Actually, from the very beginning, there were some hackathons that were set up by NCBI for different purposes, and we started there. But we realized we needed to have specific hackathons on iCn3D, and we worked with the ISMB [Intelligent Systems for Molecular Biology] meeting in 2016—the first one to actually get some new functionality in. Then we renewed these hackathons at ISMB in 2020 when the COVID-19 pandemic occurred, and we continued in 2021. That's become a yearly event, and now we are thinking about having them every six months.

Barr: That's really great. What were some of the issues in sharing reliable molecular scenes in convenient ways in the past? Can you speak about some of the proprietary systems scientists have historically used to look at these structures, and how the iCn3D model can ameliorate some of the problems?

Youkharbache: For the last 50 years, the whole community has been exchanging structural information on proteins using files—very simple files that were all stored and shared through a central database called the Protein Data Bank (PDB), which this year has its 50th birthday. All the software that was developed over the years—many different hardware platforms and software languages and so forth—were mostly proprietary, apart from maybe one or two. In order for scientists to be sharing the visualization of these files, they needed to install the software on their particular hardware platforms. Every scientist that needed to collaborate needed to have a copy of that software on their hardware to exchange files, read the files, and so forth. Very few types of software would originally be able to actually reproduce a complete visualization with all the colors and stuff.

Certainly, that would not be possible across different software. When the web came and we started having the visualization, we realized that we could share not just the files, but the complete annotations, with the coloring, the rendering, the annotation, and even the data underneath—without even having the user have any software installed or knowing anything about the files involved. That's how we started sharing visualization through web links. This works really great. These links can be put into papers. They can be embedded into HTML pages. I mean, they are very, very, very general.

Wang: These links consist of all the commands and, generally, the molecules scene. We read the data from PDB or from other sources with ID. There is a small command with ID, and then for each step you interacted through working with iCn3D, there's a command. We include all the commands in the URL, and the URL can be as long as 4,000 characters, so it satisfies most of the custom molecules scene. But if the command is much longer, then users can save the iCn3D .png image, and that .png image will have the image plus the commands or the coordinates in the file—so users can just read that file to reproduce the custom scene.

Barr: That makes more sense. That goes into our next question. What are some features of the iCn3D model for visualization, structural analysis, and sharing, like the beta plug in? Can you walk me through your design process of this resource? I know that's a very big question.

Youkharbache: I'll do the high level, and Jiyao can probably give you some more detail. It's basically an evolutionary process. We first did some visualization of proteins, and then we realized that we could actually analyze the structures, compare them, and compute some properties. That was the second part. Then the third part, which we accelerated during the pandemic, was the sharing. It was really three steps in an evolutionary process, and we have some more steps planned ahead of us. I think Jiyao can explain better, maybe.

Wang: Okay. Philippe summarized how we got from the viewer to analysis tool—and, maybe in the future, some modeling possibilities. Philippe published the paper “Twelve Elements of Visualization and Analysis.” One of the ideas is to show one dimensional (1D), two dimensional (2D), and three dimensional (3D) together in the scene, so users can either work on a 1D sequence or a 2D diagram or a 3D structure. The second point of view is that users will usually work on a subset and will usually change colors or shapes in certain subsets. Philippe mentioned that we should have a good subsets idea. Each time we use iCn3D, we work on certain subsets, and also, like we just mentioned, save the scene. We save all the coordinates and all the shapes and colors. We just save the commands, because the commands are usually small, but if you want to save a structure with all the atom properties, that could be huge. The main idea is that we synchronize 1D, 2D, and 3D, and for each operation we have a command associated with it. We start from a 3D viewer, and we add analysis tools by converting the licensed software for Delphi to RESTful API [Application Programming Interface] so users can use the viewer in 3D to show the electrostatic potential seamlessly. It's just one click and it shows potential. That's one of the examples. We convert software into RESTful API to do analysis. Also, we can calculate the symmetry for any subset of the structure. We can show the side chain prediction due to mutations. In SARS-CoV-2 there are lots of mutations, which causes the problem. We can show how those mutations change the structure and also change the interactions.

Barr: How are you integrating SARS-CoV-2 into the iCn3D platform? What work are you doing with it?

Youkharbache: Oh, that's an ongoing process. The software is general, so it can handle any molecular system. We just started working on the COVID-related proteins when the pandemic occurred. Jiyao can describe how it has evolved since then. But I want to make one point—that we keep forgetting about during the evolution—about what was key for us to be able to evolve. Jiyao mentioned the sets. The strength of iCn3D is its ability to

analyze just about anything, and its ability to select any piece of structure—in other words, at the atom level, at the residue level, at the secondary structural level—at any arbitrary subset level. It's like being able to do surgery, if you will, on molecules, by isolating the different parts. Then you can operate on them in some way—coloring is one thing, but there are many things that you can compute—in particular, the molecular interactions. That's what we pushed during the pandemic in order to study the interactions between the spike, in particular the RBD that everybody now knows—the receptor binding domain—versus the ACE2. Then we evolved the comparisons, being able to compare and do differential analysis of the interactions between SARS-CoV-1 RBD versus the ACE2, versus a new virus, SARS-CoV-2. Going forward, basically, it's actually evolving towards being able to analyze differentially what happens with the variants. That's where you can pinpoint the differences in interactions between the RBD and the ACE2 receptor. Of course, it's the same thing with the antibodies. Then you can analyze what the different variants would do in terms of the antibody binding. This is really the functionality that underlies many structural analyses of interactions, but in particular for SARS-CoV-2. Jiyao can describe how these things can evolve, especially now that we have hundreds of antibodies that are structures that are known, hundreds of structures of antibodies interacting with the RBD or different parts of the spike or other proteins. Therefore, there is now a need to be able to scale up and analyze large data sets of antibodies versus that particular spike and variants of that spike. Maybe, Jiyao, you can mention the new development that allows us to go into large data sets analysis.

Wang: Before the pandemic, we already showed annotations of SNP and ClinVar. Those are the variants in the human organs. After the pandemic, we started work on the interruption analysis, as Philippe mentioned, so we can compare different interactions. Then we introduced the side chain mutation, the mutations with the [Columbia University] scap program. There you can show the mutated structures after the point of mutation and see the change of interactions. Then, later on, we converted iCn3D to JavaScript classes so that we can write a Node.js script to batch process a large set of structures. Basically, anything you do interactively in iCn3D you can write a Node.js script for and do the analysis in the batch mode, so that's nice. This step is actually working together with the ACTIV [Accelerating COVID-19 Therapeutic Interventions and Vaccines] TRACE [Tracking Resistance and Coronavirus Evolution] project, because there we need to handle all the PDB structures within the large structure. That's where we made the conversion and made the Node.js script available. We do have some examples to calculate the epitope, to take the interactions, but basically you can write Node.js script for any functions in iCn3D.

Barr: Who uses your resource? Who is your audience or who would you like your audience to be?

Youkaribache: It's a good question. Who is the audience? It's always been historically the structural biologist and the computational structural biologists that have been using protein structures because it requires a special knowledge of the molecular structure and an understanding of these interactions in terms of physics. There is a part of physics in that—a large part, actually. Historically, it's been experts that have been trained in physics or actually in structural biology or computational structural biology. Then with the advent of large sequence databases, this field has been moving very slowly towards more experimental biologists—but very, very slowly. There has always been a need for experts that interact with biologists to actually be able to use the structure so that one can design experiments at the bench level, but there is a revolution happening right now with artificial intelligence. You might have seen these developments from both Google DeepMind and also other groups like the University of Washington, that have now actually given away their software so that everybody can use them. Also, the EBI [European Bioinformatics Institute] in Europe is now actually proposing for each and every protein sequence, a model of the structure. Now these are models, as opposed to experimental data, but they are very good for many—not for all, but for many. What's going to happen now is there's going to be really an explosion of the number of structures, with the danger of using the structure when they are not really that accurate. There

is a huge need for software now to handle large data sets, on one hand, like what we've seen with COVID-19, but also to cover the ability to use a sequence directly, and automatically produce and analyze the structure. This gives us directions for the future in order to allow biologists to use structures without the deep knowledge that's necessary to understand the molecular interactions in detail. There is a whole field that's going to open up to allow biologists to use structures in a routine way.

Barr: Do you feel that more people used your resource during COVID? Do you think COVID-19 gave you an impetus to innovate more, or faster, or that you had more attention to your resource or could do more for the future? Or is it still unclear right now?

Youkharibache: I'm sorry I might not have completely understood the question. Could you repeat?

Barr: I guess it's two questions. One question is, do you feel more people use your resource because of COVID-19?

Youkharibache: That's a good question. Jiyao, did you see something specific about the usage during the year?

Wang: Certainly, the COVID pandemic causes us to spend more time and more focus on the development, to have the new features like interaction analysis of the mutations. The usage is actually not only by the computational biologists, but also in the educational community, like online textbooks and literary texts, and some commercial companies. High school students sometimes use it for homework. You have opportunities and many functions that can let students understand better about the structures.

Youkharibache: That's one point. What we've noticed is that researchers are very often entrenched in using their own software, because they have legacy applications, and they have their files and so forth. They have their environment. They know the software—they invested time in learning about it. But for the students and the education system, first of all, the software is open source and it's free. Second, you don't need to install anything anywhere. A teacher can send an assignment and actually explain structures interactively to their students, and the students can work with the system and actually annotate the structure—and send a link to their professor to visualize what they've done. It's actually a tool to exchange between a teacher and a student and between students. As it could be between researchers, but what happens is the education system has understood that and was freed of using any software. Many of them have adopted it. What we've noticed is that during the pandemic, a lot of courses were about looking at interactions between different molecules and proteins, like, for example, remdesivir [a drug], versus its target, the proteases, or between the RBD and the ACE2 protein. Actually, it was used for teaching more during the pandemic, but people will use it on an ongoing basis on any system.

Barr: That is really interesting. Has that influenced your design at all, knowing that it's being used in schools?

Youkharibache: I don't think so. I think it got us to understand that this is very powerful, and we need to push it much further. I mean, there are many applications we've thought about. These links could be there. One of the problems is many of the links disappear after a few years, because the website disappears and so forth. But here the idea is that, since it is developed by NCBI, it actually will be maintained by NCBI, at least for the time being and for the foreseeable future. The other thing is, we've plugged a lot of different things into the software, so there are actually date stamps for all these annotations, and the links can be reused seamlessly five years from now, for example. Even if the software evolves, the previous version of the software can be used systematically. And all that is available on an open-source manner, actually, on GitHub.

Barr: That's really nice. You've mentioned some of them, but can you go more into detail about some of the challenges that you and others on the team have faced with developing the resource?

Youkharibache: I'm sure Jiyao can say a lot about the challenges.

Youkharibache: Oh, yeah, it's challenging because once your software gets complicated, it has many functions you want to maintain during each release. It's challenging because there are so many inputs and so many operations where you can change the properties and so many ways to output. Also, we need keep it backward compatible with the previous shareable links. It's challenging, but we try our best. We also introduce unique things to iCn3D, like showing the electron set potential seamlessly, doing batch analysis, or sharing links. I think these are the unique things about iCn3D. People can use their previous software, but if they want to try something easier or something with new features, they can try iCn3D as well.

Youkharibache: I would say that it's challenging, but actually Jiyao has taken on this challenge, really, because there is so much functionality today, that maintaining it is a huge job. We both have to maintain it and develop new features to go into easier applications and automatic analysis. The truth is now it's becoming really a resource issue, so the problem is twofold. One [problem] is that we need to develop a community of developers behind that open-source software, because we cannot do it all. And then the problem is also for NIH to realize that we need more resources. That's the two parts of the puzzle here going forward.

Barr: On that note, what are some next steps for the iCn3D resource? One of the issues I read in your abstract was numbering, which has been a difficulty in the past, but is something that you're working on. What are some other hopes you have for the resource?

Youkharibache: Yeah, numbering is an ongoing issue. In the last week, I think Jiyao and the NCBI team has been working on this problem. Again, it's a very challenging problem.

Barr: Can you explain why it's so challenging? Sometimes people may not understand that.

Youkharibache: Do you want to explain that? It’s so fresh in your memory now, Jiyao.

Youkharibache: In PDB structures, sometimes they want to give the same residue number to certain residues with certain functions. Sometimes they have 30 residue numbers or maybe 100. Then they also want to define as 100, so they add a letter update—so 100-A, and another residue is 100-B. Typically the residue numbers are continuous from one to two to three and going forward. And if you add a letter to a residue number, it makes it not an integer anymore, and it's not continuous. They introduced all kinds of problems.

Youkharibache: That's one issue, and actually it's particularly acute in the antibody world. And antibodies are everywhere now. This probably existed since the early days of a guy named [Elvin] Kabat, who developed his own numbering for antibodies. These things grew, of course, with complexity, because we discovered different properties of antibodies in terms of length. The numbers had to be played with by keeping the same number but adding numbers where they couldn't be sequential numbers anymore. But you have that in COVID-19, for example. If you're going to take SARS-CoV-1—which we did in that preprint you're referring to—versus SARS-CoV-2, well, even the RBD don't have the same numbers. Therefore, when you say, “Well, there is no variance that has actually changed this residue at position 484.” Well, that's not necessarily 484 in all the files, and certainly not in other related viruses. All these things are actually of paramount importance although they are

very mundane things, but we have a lot of work with that, which takes a lot of time. Now, this is one problem, and I think you asked for more than just this challenge, right?

Barr: Yes, or other things that you are planning on doing as you move forward.

Youkharibache: Jiyao could tell you how much pressure there is now to be able to analyze not just the PDB structures that have been developed experimentally, but all these models. We went from having 150,000 or 170,000 structures to now having millions, and we're going to have way more than that, because this structure is covering approximately 18% of the human proteome. Now what has happened is that the whole UniProt database for humans is covered, and it's covering about 98% of the human proteome. They're going to add 20 more proteomes, so you're going to see the same explosion in structures that you've seen in the human versus other genomes, in terms of proteomes, with even bigger complexity because the genome has a very specific strength. It had a numbering that's based on the human genome, and that numbering actually has a particular position. Now, proteins are all over the place if they can form assemblies, so you have not only a proteome, you have all the interactions that can exist between these proteins, so that will present millions of structures to analyze down the road. The challenge for us is to bring in the ability to handle these data sets, but also to allow biologists to understand these systems. That requires a level of software development where we plug in a lot of intelligence in the system, and ease of use, which the field is not accustomed to. The experts have to actually become something like "cognititians" used to be, where we understand what is needed for biologists to use this information to understand the biology in terms of atom interactions. This is a huge challenge. I think this is probably the one of the biggest challenges.

Barr: That will be very interesting and very difficult to handle.

[laughter]

Barr: That's an understatement, I guess. You all are laughing!

Youkharibache: I think we can say it's an understatement when you realize the resources we have.

Barr: Yes. How have you all promoted this resource for scientists all over the world to use, and have you received any feedback from those who have used it?

Youkharibache: Well, Jiyao might have more feedback than I do, right?

Wang: Yes. We have several headphones, so we interact with users directly and get their feedback. When we develop new features, we can get the feedback. We have communication with previous iCn3D users that are familiar with iCn3D, so they know some of the features. That's another way to set up another community. We also communicate with teachers of the online textbook library texts, and also some teaching communities. Some users send a problem to the Help Desk. Some users communicate with us directly because they have more involvement with iCn3D.

Barr: Can you talk a little bit about what each of your roles have been? It's a little clear that Jiyao has done a lot of the development, and Philippe has done a lot of the big ideas, but can you talk just a little bit more about what you all each have contributed and some of the other people on the team and the expertise that they've added to this project?

Youkharibache: I think Jiyao can tell you about how things happened. What has happened is that I've basically used a software with real needs, and expressed these needs in terms of functionalities that we should have, as well as some design elements. Then from there, Jiyao has engineered—and re-engineered—the software constantly, although also using the infrastructure at NCBI. I will let him talk about that.

Wang: Okay, so it's actually my previous advisor, Steve Bryant. He's the group lead. He initiated this project to use iCn3D. At that time, Philippe was also at NCBI, so he helped design this software. The Structured CDD [Conserved Domain Database] team at NCBI provides all the back-end data, because NCBI does annotate the PDB structures. Then from the back-end, they provide all the data, and also all the VAST alignment data. Actually, Philippe and I worked together on new features and how to design and implement them. He also tested them because he's the first user for most of the new features. He finds bugs. I also sometimes communicate with other developers for other software like NGL Viewer or 3Dmol. They also provide very valuable feedback for some new features. For example, NGL displays the atoms or cylinders or spheres used in imposters that will speed up the display. And from 3Dmol, we show how to generate the surface. We share the source code among different software, and we help each other. In a way, I see it as kind of another community. Even though we develop different software, the ideas are the same. Most importantly, it's about annotations. NCBI has the Conserved Domain annotation, the VAST+ alignment annotations, SNP annotations, ClinVar annotations. And EMBL-EBI provided AlphaFold structures. As a community, we provide all kinds of annotations and features. It's really a community effort.

Youkharibache: I should say that this is actually an interesting point. The strength of NCBI developments that have been occurring in the last 20 years is in the annotation of these structures. What's happening now is this explosion of structures all over the place with the fact that they are just models. The need to annotate these models in terms of their validity and invalidity is something that actually alpha 4 does to some extent. But when you want to relate to other structures, you can actually do a lot more annotations. When you relate to sequence database, you do a lot more annotation. We have all these annotation systems at NCBI that now need to be scaled up and provided to annotate the whole world of structures, and this is a huge challenge that has not been addressed yet. But we need to transform the architecture of the annotation servers so that they can be valid for the whole world, opening up the system if you want. That is actually a very, very crucial thing, because, yes, right now there's another aspect—the aspect that Jiyao was mentioning about the community that exchanges software. It exchanges software, but this software needs also to be organized as shareable libraries and APIs, so that new applications can actually be developed based on these elementary modules. This is also another challenge for the community—to come together and develop these libraries and this API—something that everybody's been talking about, but that's also something that needs to be done at the community level.

Barr: Many, many things to do. Have you all worked on any other COVID-19 research or taken part in any other COVID-19 initiatives? I know you said you worked a little bit with ACTIV and TRACE. Can you talk more about that?

Wang: In the ACTIV TRACE project, our CDD group is mainly focused on annotating the 3D structures and providing the Conserved Domain information. We provide the epitope information, the interactions, and we can also provide the 3D visualization, if people are interested. I mean, other people are working on the strings and where the mutations are. We also collaborate with IEDB [Immune Epitope Database & Tools]. They provide annotated epitope information. We also do our own analysis about epitopes.

Barr: How much time do you get to work on your iCn3D resource? I imagine you both have other projects assigned to you in your job, and it seems like iCn3D could take all day, every day, if you wanted it to. How do you allocate your time?

Wang: That's a very good question. Starting from the beginning, I think I was pretty much full time for about two years, probably. Then afterward, I spent half of my time on it and spend the rest working on other things. Right now, we convert to agile platform. Everything has to be reported and has to be assigned, so I get a less and less resources to work on iCn3D. At most, one third to half my time is spent working on iCn3D.

Youkharibache: I work on the iCn3D on the application side. It's a tool that's always available. Like you have a word processor when you type, well, I have iCn3D when I look at molecules or when I analyze molecules on different projects. Of course, it's like the early days of word processing. You need new functionality that you don't have. Every single problem brings a new need for a new functionality. That's when I started working with Jiyao, designing a new functionality. It's an ongoing process all the time. For me, it's part of the whole work that I do on the application side, that leads to asking a scientific question about an interaction, for example, and then needing to see that interaction, to actually quantify that interaction, and to compare that interaction. All that leads to a need in terms of new NFEs—what's called “new features and enhancements.” This is done in an interactive way with Jiyao, who, when he can, gets to it. If I try to present things as a bug, it goes faster. We use a number of tricks to make things happen.

Barr: Is there anything else that either of you would like to share about iCn3D or about your work with COVID-19, or your experience during COVID-19 at NIH?

Wang: I want to mention some things which are unique to iCn3D, like that you can write Node.js script and use some examples to do batch processes on large sets of structures. That's something I think is unique to iCn3D. You can share the link with your colleagues, or you can even show it in a publication if you want. It’s an interactive viewer in a publication. That's one possibility. We converted several software to RESTful API, to use 3D viewers. Basically, most of the computational groups, when they provide new software, provide a website, and some of them provide the RESTful API and some don't. The RESTful APIs are really useful and crucial to share the resource with the community.

Youkharibache: The thing to share is about where we want to go. We want to be able to analyze large data sets so that we get information about the biology of a system. There is the analysis part, but there is also a design part. Once you know what you want, you need to be able to actually model structures and then have experiments that relate to these models and so forth. Really, it's an evolution in different directions in depth. That's what we've always done. We were going to analyze in depth, but also now we want to analyze on scale, and we want to be able to use that information. The word here is “information”—to actually design new systems and new drugs, and not just drugs like small molecule drugs, but new proteins, new chimeric antigen receptors, and new chimeric proteins of any kind that can interact with systems in biological systems in disease states. That's really the challenge—going closer and closer to the biology and to the pathology and the clinic. That is where this needs to go. The structure is one element of understanding—a huge element of understanding—this system. That's where we want to go.

Barr: Well, I wish both of you the best of luck with everything and thank you very much for speaking with me. I hope you and your families continue to stay safe during COVID.

Wang: Thank you for having me.

Youharibache: Thank you very much, and the same to you. And thank you for taking the time for this interview, because we don't have many opportunities like this to explain in detail what happened. Thank you very much.

Wang: Thank you.