Dr. Nadya Tarasova
Behind the Mask
August 27, 2021

Barr: Good afternoon. Today is August 27, 2021. My name is Gabrielle Barr, and I am the archivist at the Office of NIH History and Stetten Museum, and today I have the pleasure of speaking with Dr. Nadya Tarasova. Dr. Tarasova is a senior associate scientist in the Laboratory of Cancer Immunometabolism at the National Cancer Institute, and today she is going to be speaking about some of her COVID research using the Biowulf supercomputer, as well as some of her other pandemic experiences. Thank you for being with me.

Tarasova: Thanks for having me.

Barr: To begin with, what is a Biowulf supercomputer, what are its capabilities, and why was this technology necessary to conduct your COVID-19 research?

Tarasova: Biowulf is simply a cluster of over 95,000 cores, basically 95,000 computers that operate in Linux platform. This is one of the largest in the country, and it is designed to run simultaneously a large number of jobs, which is exactly what we need in drug discovery because we all thought that one day the drug discovery will become computational.

Barr: That is an interesting thought.

Tarasova: Right. We all kind of believed in that although there were lots of skeptics, and there are still skeptics. I used to be one of them also.

Barr: That is interesting. Why were you skeptical before, and what changed your mind about it?

Tarasova: Because it did not work. We simply did not have the tools, but a lot has changed. One of the biggest components that changed was the computational capabilities that we did not have, but also the tools were not working well. [However], algorithms have been improving constantly; they have improved so much that, as a matter of fact, nowadays, virtual predictions work really well, which means that instead of running very expensive wet experiments, you do it all in silico [performed on computer or via computer simulation].

Barr: Have you used Biowulf before the pandemic?

Tarasova: Yes, because we started virtual drug discovery about three years ago, probably a little bit more, because other very important components of computational drug discovery are these virtual libraries of compounds. That is part of what our work is, and now you know I am going to pepper you with some numbers.

Barr: Okay.

Tarasova: These are, believe me, important numbers because the chemical universe is very large. It is estimated to contain 10 to the power of 63 of drug-like molecules.

Barr: Wow!

Tarasova: Yes. PubChem has just 100 million described, and these are the ones that were ever made. So, just 100 million compared to—that is 10 to the power of 8—this is like invisible in this universe. That is so small. And definitely, we want to look at the entire universe to be effective, but also, we do not want to look into the compounds that are kind of theoretical structures; we need them to be synthesized. That's another big problem—huge, very challenging—is to predict which structures can be made and which just might exist on paper, and they may look stable, but nobody can make them, so what is the point looking into them? That is part of the work we do. We generate those libraries of synthetically accessible compounds.

Barr: It is really interesting.

Tarasova: Yes. That is a lot, but this is all recent and as I mentioned, we also [anticipate] that one day drug discovery will be computational, but guess what? That day is now!

Barr: Yes.

Tarasova: It is here. Not everybody realizes that because those tools are not widely accessible for several reasons. One of the reasons is computational power. We are very likely at an age to have viable power.

Barr: Is it just very expensive to run those kinds of tools, or do you need a lot of space, or what are the constraints of why it is not widely accessible?

Tarasova: Yes. Maintaining Biowulf cost lots of money. You can do those types of searches that we do on Biowulf, and you can do them in cloud. We are now just starting a collaboration with Google because we can expand beyond capabilities of Biowulf, but it costs lots of money. We are talking about really big, big, big money. It is hundreds of thousands of dollars a month, [or even] a week.

Barr: All right, that is a lot. How does it, well, it is maybe too early to predict, but how does it compare to the money spent on physical drug trials because they are also costly?

Tarasova: It is, also, but it is definitely still much, much, much, cheaper.

Barr: Okay.

Tarasova: But the major advantage is that by experimental tools you cannot screen that many compounds. Currently, realistically, a top-of-the-line screen is about three million compounds. That is the best, the largest, the most diverse libraries that is several millions. When we do virtual, our current library, which is called SAVI, that Synthetically Accessible Virtual Inventory, is 1.7 billion.

Barr: I cannot imagine.

Tarasova: Yeah, but this is important because it increases your chances of success.

Barr: Yes.

Tarasova: Since I have been working in cancer drug discovery a lot, one of the major challenges in finding good treatments for cancer is that lots of proteins that we need to shut down lots of bad guys, mutated ones. They are what is called “not druggable”, and lots of them have been known for a very long time, but simply, lots of attempts to find compounds that shut down their functions were in vain. I think that was partially because we were searching in a really small part of the chemical universe.

Barr: Yes.

Tarasova: Now, since we are expanding this part of the universe, we have started finding drugs or inhibitors for proteins that we could not target before.

Barr: With Biowulf, I know that it creates simulations, and we may get to this later, but can you do things like put in a filter and show how the variants affect whatever protein, such as the spike protein, and then find the drug that would be better for how the cells react in that situation, or put in sort of situational things?

Tarasova: Yes, it is quite possible. There is one limitation already to the virtual drug discovery, however, because we need to know protein structure, and if there is a new variant, we need the structure of the variant. Sometimes you can easily simulate it, but the more precise the structure is, the more precise are the predictions from the virtual screens. Luckily, there is a silver lining to everything, and there was a silver lining to the pandemic as well. We were not allowed to work on anything [NIH laboratories switched to COVID work], so these scientific resources were thrown in this direction, and the structures of SARS-CoV-2 proteins have been coming in an avalanche, lots of them. For me, it was a challenge because crystal structures sometimes represent solution structures not quite accurately. There is also another challenge that currently we cannot tell looking at the structure whether it corresponds well to what is in the solution, or it was twisted a little bit during crystallization. That is why we had to test lots of structures.

Barr: That makes sense.

Tarasova: Some of them did not [correspond] so, we had to synthesize lots of compounds just to choose the structures that provided good predictions, but generally speaking, you need the structure. In the past, it could have been a bottleneck, but not as much nowadays because crystallization procedures have improved. The methods of structure determinations have improved and became more effective. And now we are getting into a new era of determination of protein structures without basically any experimentation. A new database has been released three weeks ago, no, a little bit longer, that contains predicted structures of over 300,000 proteins that are determined by a software called Alpha Fold.

Barr: So, you can be ahead of the curve now?

Tarasova: We hope so. We do not know whether they will work. Although I’ll tell you the secret is, of course, we are already testing whether those predicted structures allow for good predictions, but we are in a total new era.

Barr: That is very exciting.

Tarasova: It is very exciting. It is very exciting.

Barr: When did you begin searching for SARS-CoV-2 protein or ligands in the new ultra-large libraries and its synthetically accessible compounds, and can you also say how extensive these libraries that you look through to find these ligands for SARS-CoV-2?

Tarasova: Yes, the libraries were quite a story of their own because, as I have mentioned, it is very difficult to predict whether a compound can be synthesized or not because you need all this chemical , organic chemistry, knowledge that has been created. And here is the challenge: when you have a reaction, you have to know whether it will go, or it will not, and you will get a huge mess. The only way to predict is based on prior knowledge, prior results, but the problem is that nobody publishes negative results, and that makes this prediction quite challenging. Savvy Creations is an international collaboration based on a very old algorithm for which Elias Corey got a Nobel prize in 1990. It was an algorithm created at Harvard University a long time ago, and we are still using this special computer language that was written for that, that we still use. One of our major collaborators, Philip Judson in U.K (United Kingdom), continues developing it for us because the original script of LHASA (Logic and Heuristics Applied to Synthetic Analysis) had a very limited number of reactions, and some of them are just dated reactions, chemistry that we do not use. One of my roles here on the project is to write new chemistries for Philip, who translates them into a computer language, which is a program. Then, a group in Germany translates it in a language that is called Cactus, can do virtual synthesis, after which the group of Marc Niklaus at NCI (National Cancer Institute) does this virtual synthesis on Biowulf. This is kind of a circle; then again, my group uses the generated libraries to test them, to verify that they are working, and we synthesize them. Currently, the synthesis is conducted mostly in the Ukraine. We have a collaboration with company called Enamine, and SAVI has been created using their synthetic blocks, the ones that they have on the shelf and have been verified. The success of the synthesis was predicted correctly in 95% of cases, which is really high.

Barr: Can you talk about your process of searching through these libraries and how you filter? Did you have to do anything—you talked a little bit about the algorithm, but did you tweak the algorithm? Any other kind of algorithms to get things?

Tarasova: Yes, we are developing them all the time, and here is another computational big challenge because there are two approaches in general in virtual drug discovery, in searching the libraries. You can do what is called the receptor-based search. You have a pocket on the structure of the protein, and you are trying to fit different molecules into this pocket, but here is the problem: you have to try lots of different conformations of the molecules for each molecule with lots of different positions inside this pocket. This is a relatively slow process that takes about 30 seconds per compound. It looks fast, but if we would be searching the entire cyber database by this methodology, even with the parallel processes, and we can run up to 2000 in parallel, currently, it would take almost 300 days to do this.

That's a lot. This really demonstrates that although we have really one of the best and biggest computers in the world, it is still small and this is still a challenge, but we all know that computational power grows logarithmically, so there is light at the end of the tunnel we can see. We will be able to do more, but right now we have to go around this problem, so going right to the heart of the problem, what we are working on is developing methods that access of all these databases much faster. We do what is called legal based screens; once you identify your hits from smaller libraries, you, then, look for similar molecules. We are just looking at the conformations. And here we also collaborate a lot. I mentioned our collaboration that we have with Google, where we can compare 3D molecules, preferably looking in spatial 3D arrangements. We also collaborate with another company called

MolSoft that is based in San Diego, that is really, kind of a leader in the field in terms of developing software for the streams, and it was a challenge to run this even in Biowulf because although Biowulf is wonderful, it has limitations. For instance, the largest disk that we can get on Biowulf is 3.2 terabytes and the library of all conformers for cyber is 4.8 terabytes.

Barr: Okay.

Tarasova: It will not fit. It is just huge volumes, but again, this is an archive. Very soon, we will be laughing at this limitation.

Barr: Yes.

Tarasova: For now, 4.8 terabytes seem a lot, but it will not seem so in the future. The same thing, as I remember, when 250 megabytes seemed like a lot.

Barr: Can you talk a little bit about what the simulations look like when they are produced on Biowulf?

Tarasova: When you run the software, it is beautiful. You see all the proteins in three dimensions, but then you submit the job. You do it on Linux, and that is not beautiful at all. It is all black screen and lines, lines, lines. There is a lot of programming and writing scripts involved because you have to run those screens in parallel, and there is a line for every job.

Barr: How do you know if something will work? How do you see that, or do you see it in the script, or do you see it in the simulation, and are they colored or are there lines? How do you know that your drug will work against whatever it is?

Tarasova: Basically, the software simply screens and throws away what does not fit, and you end up with what the software thinks will work. Usually, we conduct screens and rescreens because although a slow screen is a slow screen, nevertheless it is faster than the synthesis, so I would rather screen and rescreen than spend lots of money and time on purifying and synthesizing the compounds. But still there is always a risk that, for instance, the structure is not quite accurate, or something else is wrong. In our experience, usually, it is all or nothing. If the structure is correct, then all predicted compounds will work; if structure is incorrect, then nothing works. That is why we do those quick screens, using smaller libraries, small, I mean about three million compounds.

Barr: That makes sense. Can you talk a little bit about what your findings have been so far in terms of finding possible drug candidates for SARS-CoV-2?

Tarasova: With SARS-CoV-2, another challenge, of course, is testing compounds. We do lots of wet experiments in the group so we have verified bindings using biophysical methods, and again, so we are lucky because NCI (National Cancer Institute) has biophysical resources that have lots of cutting-edge technologies that we can use for that. And here there are a lot of challenges that we have gotten. There are active binders to quite a number of SARS proteins, but they have been tested on a live virus, so this is not an accurate prediction because we are predicting only binding, but this is the biology of the virus [which is showing] what is going to impact the biology and what is not going to impact. Originally, the only method that we have enhanced now is viral entry, and in viral entry, only one of the classes of the compounds, and not surprisingly target spike protein. But what is unusual is that those compounds bind in a very unusual pocket, which we call allosteric pockets, and this is not where H2 binds, for instance.

Barr: Can you talk more about that? That is really interesting.

Tarasova: It is very interesting because the function of the spike is quite complex. It undergoes multiple conformational changes. There are lots of hinges and this compound sort of blocks the hinge, blocks the movement, that is what we believe it does. The nice thing about that is that it is highly conserved, so it is far away from the mutations that occurred in variance. We are still pretty early in this thing so, because we have the data from the first round, now we are in the process of the optimizing the compound. The bottleneck is the assay.

Barr: Yes.

Tarasova: It is very expensive, as you can imagine. It is a live virus.

Barr: Yes.

Tarasova: It is done currently in the National Center of Advanced Translational Science.

Barr: Have you looked at drugs that target the N and the M protein?

Tarasova: We, on purpose, did not try to do anything with these proteins, for instance, because industry put such humongous efforts in that. I thought we would better do something that is what the NIH is supposed to do, to do something that nobody else does. We also have some binders to end protein, but sadly, they did not impact, or at least, we did not see an effect on a live virus. We know that they work in vitro, in the nucleocapsid protein itself. We are currently working on the nuclease, and the problem with the nuclease is that there are no good assays, biological assays, because nuclease inactivates the immune system, and that makes the virus invisible for the immune system, and that is very difficult to assay. We have potent inhibitors in vitro activity but I have not tested that. Also, transcriptase is interesting, but the challenge with transcriptase is that we do not have x-ray structures.

Barr: Okay.

Tarasova: Only cryo-EM [Cryogenic electron microscopy] structures, but generally they do not have the resolution that looks as well as x-ray. But we are trying to see, because the best available resolution, which is 2.5 angstrom, and that comes close to x-ray. We have some leads but we need to test them on the virus.

Barr: Do you need both the cryo-EM structure and the x-ray structure to most effectively do your part of the work?

Tarasova: No. Usually, x-ray is enough.

Barr: Okay.

Tarasova: But simply, crystallization turned out to be very challenging for the transcriptase, so that is why nobody has the structure that we need. For cryo-EM you do not need to grow crystals. Because everything is done in solution. And cryo-EM is developing at an amazing speed, there are already reports of new hardware that allows for a resolution that gets very close to the best x-ray, but this is definitely

not very accessible because those electronic microscopes cost an enormous amount. We are talking about several million dollars.

Barr: Wow! That is more than I thought it was going to be. Can you talk about how your earlier work looking at SARS-CoV-2 proteins has spawned any further or new COVID research for you? Are you continuing on looking at structures and drugs?

Tarasova: We continue on because it seems to be working in general, but since we are back to the lab now, I also have to get back to cancer research.

Barr: Yes.

Tarasova: Indeed, it is unfair to cancer patients. They need our attention. But we have learned a lot doing COVID work, and the tools that we have been developing in parallel work for any disease.

Barr: Right, yes.

Tarasova: That is another part of the silver lining, and I think that my group became more productive during the pandemic.

Barr: Really?

Tarasova: Yes.

Barr: Why do you think that the pandemic made them more productive? Or how did the pandemic make them more productive?

Tarasova: Because we were focused, although we were probably exhausting ourselves because I would be working frequently 18 hours a day, submitting jobs when I got free nodes available—sometimes in the middle of the night—because everybody was in a rush. Sitting at home, I had all the computational capabilities I have in the lab—the same speed and everything else—but in a normal situation at work, I would not be operating [that way].

Barr: Right.

Tarasova: And on the weekends. That is why we became more productive. There was also this sense of urgency.

Barr: How has it been conducting cancer research during the pandemic? Can you talk a little bit about that?

Tarasova: We were not allowed to do any cancer related experiments in the lab, and unfortunately a project we were involved in had to be put on the back burner, just because we were allowed to do COVID research and that focus actually took our entire efforts. Now we are back at work and COVID testing is done mostly at NCATS.

Barr: Yes.

Tarasova: Which is nice, while we focus a lot on old cancer problems. But again, that benefited because during this COVID pandemic we were optimizing the libraries our research did not slow down.

Barr: Right

Tarasova: It accelerated, as a matter of fact, because there was more of a sense of urgency, everybody was very productive, and it is noted that if you are doing computational research, you are more productive at home.

Barr: Yes. I have to ask you. Were you always interested, from your earliest days as a scientist in data because it seems like you really work with that right now, with a lot of computational elements of science?

Tarasova: I really liked the precise part of science. I was pretty good in math since my earlier days. As a matter of fact, I was considering during my entire childhood a career in engineering, but at that time, it was an old boys clubs. I was advised to choose a field that was more woman friendly. Now things are changing, and I am happy to see that. I grew up in Russia, and I was the fourth in physics in the entire Soviet Union, but my principal advised me not to choose the career in physics, but I always loved doing things that are very logical. I never regretted that I went into chemistry because chemistry comes in so many flavors, so that is what I am doing, but the important thing is that it is very meaningful.

Barr: It is very meaningful and it has a big impact on people. Do you have anything else that you would like to share about either your COVID work or experiences both as a scientist but also as a person who has been living through this pandemic like everybody else?

Tarasova: Like all scientists probably, I am very upset with people who ignore science. That is my personal attitude, and I am also very frustrated about the division in the society that we have to live through. I am sure that we will overcome this adventure because any society is always going to change.

Barr: Yes.

Tarasova: Nevertheless, as a society we keep improving.

Barr: Yes. Well, you have a very positive outlook, which is really good. I wish you and everyone in your lab continued success. I definitely hope some of the drugs that you are screening work in real life because we need them, and I look forward to seeing what else you and your group do. Thank you.

Tarasova: Thank you. It was a pleasure talking to you Gabrielle.