The Dark Side

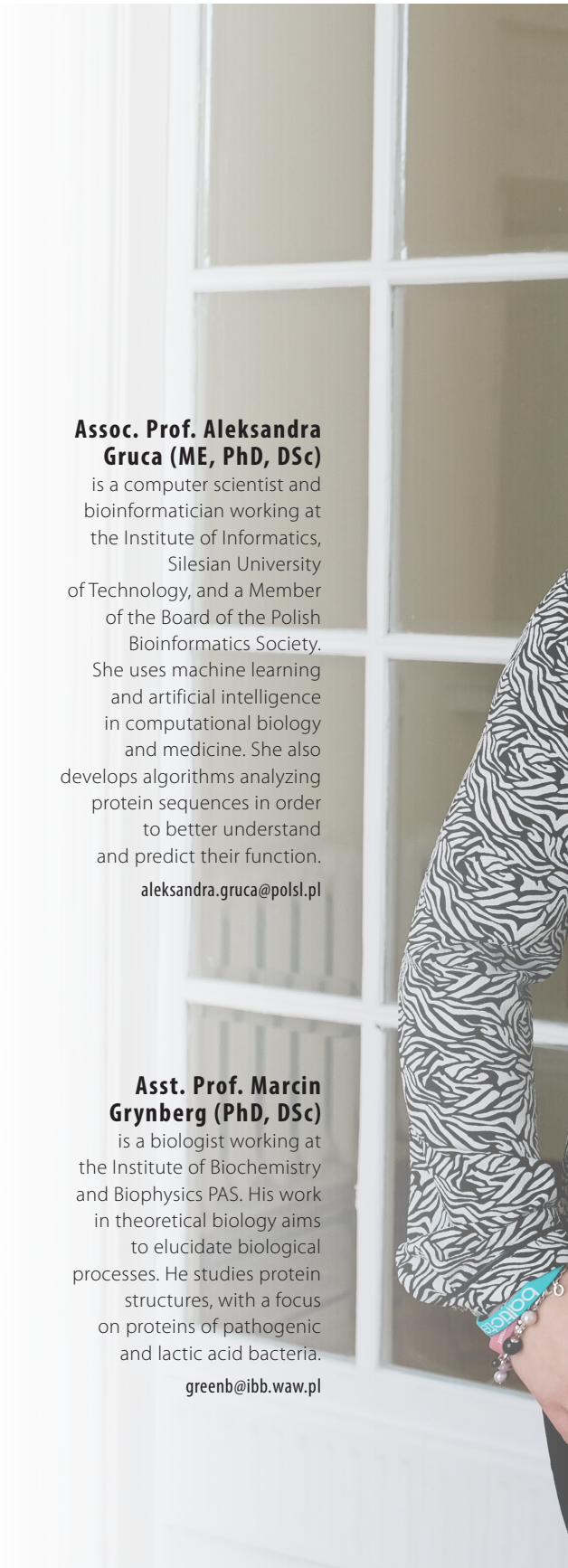
They are unfairly regarded as junk protein, but the truth is they have not been properly studied due to a lack of suitable tools. So how much do we really know about non-globular proteins?

ACADEMIA: **What are non-globular proteins and how do we study them?**

MARCIN GRYNBERG: Globular proteins are long molecules which arrange themselves in space in a certain way – they fold into a roughly spherical shape. Non-globular proteins do not form such structures; instead they exist floating freely within cells. The traditional approach to studying proteins requires obtaining crystals of a large quantity of the given protein, which generally can only be grown from proteins with a regular structure, i.e. globular proteins. The process is so difficult for non-globular proteins as to be almost impossible.

An interesting property of non-globular proteins is that their sequences of amino acids – their building blocks – tend to be completely different than in globular proteins. As well as being difficult to image, this property has led to their largely being ignored in research for decades.

ALEKSANDRA GRUCA: The topic has also been avoided because researchers originally believed that protein fragments comprising low complexity sequences are not important. They were regarded as junk; today the approach has switched to discussing the dark proteome, describing the protein world which had not been studied due to a lack of appropriate tools.



Assoc. Prof. Aleksandra Gruca (ME, PhD, DSc) is a computer scientist and bioinformatician working at the Institute of Informatics, Silesian University of Technology, and a Member of the Board of the Polish Bioinformatics Society. She uses machine learning and artificial intelligence in computational biology and medicine. She also develops algorithms analyzing protein sequences in order to better understand and predict their function.

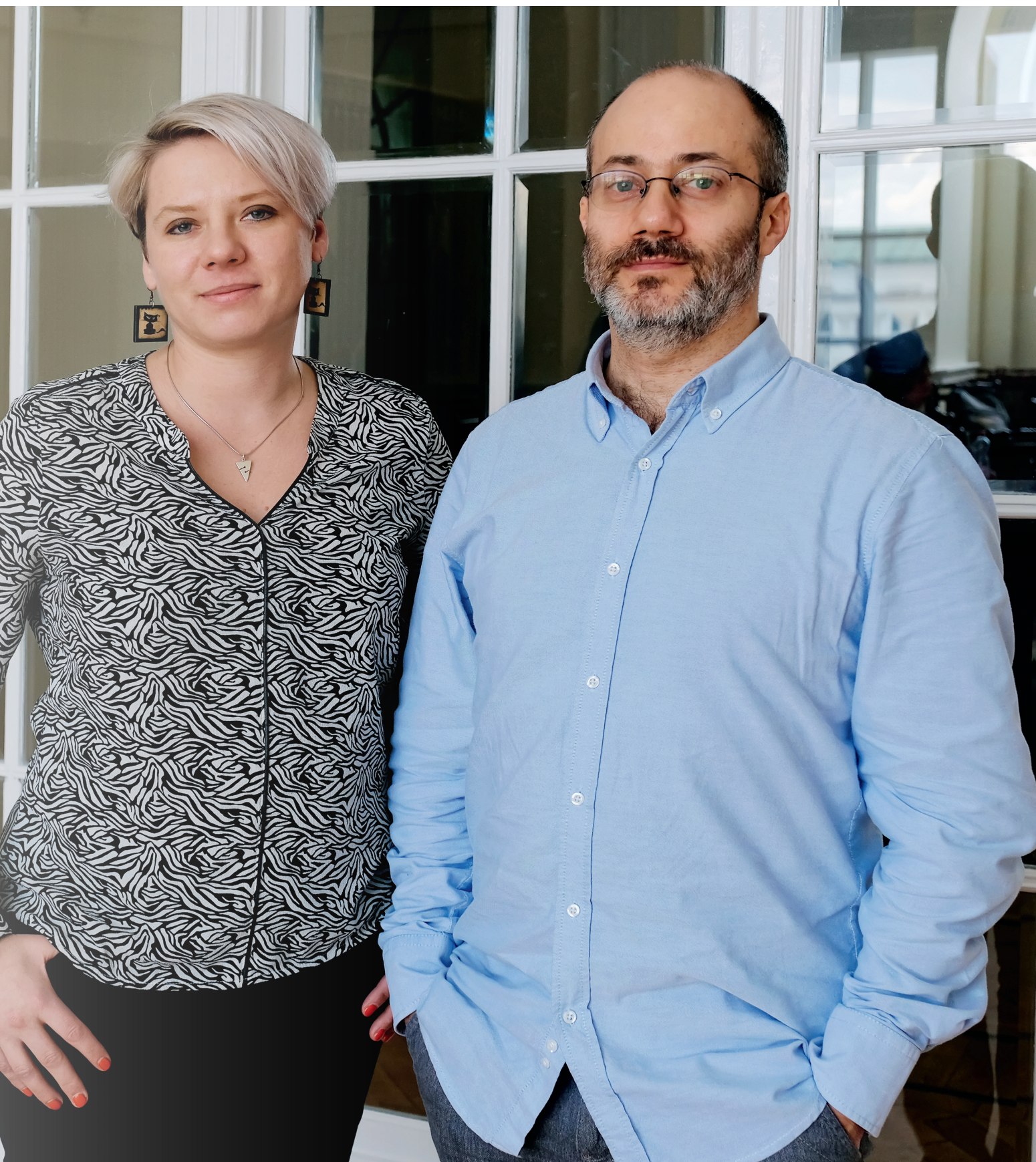
aleksandra.gruca@polsl.pl

Asst. Prof. Marcin Grynberg (PhD, DSc) is a biologist working at the Institute of Biochemistry and Biophysics PAS. His work in theoretical biology aims to elucidate biological processes. He studies protein structures, with a focus on proteins of pathogenic and lactic acid bacteria.

greenb@ibb.waw.pl

ASSOC. PROF. ALEKSANDRA GRUCA, ASST. PROF. MARCIN GRYNBERG

OF PROTEOMICS



You are both using the past tense – what's changed so that these proteins can now be studied?

M.G.: Researchers have been trying to study these non-globular fragments for a long time, but with limited success. The crystallographic methods I mentioned earlier try to capture proteins in a "frozen" state. It's as though I were walking down the street, suddenly froze on the spot, and someone snapped a picture and said "Right, this is what Marcin looks like." And while that would be true, I can also look a myriad different ways. The problem is that crystallography does not provide an image of protein fragments while they are in motion.

Nuclear magnetic resonance (NMR) visualizes moving objects, but it cannot be used for large molecules. Only very short proteins can be studied. This raises the question: how do we place these short fragments in the context of a larger structure?

Another reason why studying non-globular proteins is difficult is their sequence composition. Globular proteins are built from 20 amino acid types, and a "normal" protein contains most of them with a certain regularity. However, non-globular proteins frequently include fragments with far fewer types of components, comprising three or perhaps five repeating amino acids. We call these low-complexity fragments. From a mathematical perspective, analysis of such fragments is difficult because we have very little information to go on. Although it might sound absurd, this is in fact the case: if a fragment is repeated, it means there's little information and that makes the protein all the more difficult to study mathematically.

> Some proteins are harmless individually, but if they combine they can create toxic structures.

A.G.: Statistical models which allow us to analyze similarities between protein sequences (which we can use, for instance, to predict protein structures) have been developed for fragments which contain all 20 amino acids. The models have been in use for over 30 years, but they simply cannot be used to analyze low-complexity regions of protein sequences. Because tools to analyze these fragments did not exist in the past, they were simply excluded from analysis, with researchers instead focusing on protein fragments comprising all the amino acids.

Do we know the function of non-globular proteins or low-complexity fragments?

M.G.: Because of their unusual structure and sequence, many of them are involved in transporting other proteins between cellular compartments, or in converting them from the liquid to the solid phase. They also frequently bind proteins to one another, or to other structures within cells. We are aware of a few other functions, but all this is still a drop in the ocean.

A.G.: Some of these proteins are harmless when they exist individually, but if they combine they can create structures which are toxic to the organism.

M.G.: That's right – a protein can precipitate out of a solution and form a polymer, like for example specific proteins in certain cancers. But non-globular proteins can also have other functions. Many are involved in binding other proteins or signal-transmitting molecules. Research done over the last decade shows that they can play a part in creating new functions. Because non-globular proteins or their fragments can be found in many organisms, they are easy to trace. It turns out that at times low-complexity protein sequence fragments can evolve faster than others, going as far as developing new abilities such as turning into enzymes performing a new function.

A.G.: Because these proteins have a non-defined structure, it makes it easier for them to encounter other proteins and create something new with them. Think of it like this: if someone is curled up in a rigid ball, it's difficult for them to interact with others, but if they're waving their arms around, there's a high likelihood that they'll catch someone else by chance.

M.G.: Or even not by chance.

You recently organized a hackathon as part of the COST consortium.

A.G.: COST stands for Cooperation in Science and Technology, and it is Europe's longest running research program financing meetings between scientists from different countries working in the same areas. It currently has 36 member countries, including all EU member states and associated states.

M.G.: Our own consortium is called NGP-net and we first came across it via Silvio Tosatto, its founder from the University of Padua. Silvio was looking for partners in Europe and approached my colleague Anna Gambin, a professor at the Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw. She was too busy to get involved, but she put Silvio in touch with me, and I later found Aleksandra. It was really by chance, because I first came across a program written by Aleksandra which interested me so much I asked her for help and involvement, and luckily she was interested.

How do mathematics, computer science, and modelling help in researching proteins?
A.G.: At least half of the members of NGP-net are bioinformaticians – researchers who study proteins without ever setting foot in a biology or chemistry lab. We sit in front of computer screens looking at structures and sequences, and try to make sense of it all.

During a classic hackathon, programmers spend a few days working hard to develop new software or a prototype. What about your protein hackathon held last February?
M.G.: Our meeting was short, so we focused on planning what the final product should look like. We published an article on the tools and terminology concerning the subject, since we realized there is a problem with definitions which needed sorting out.

Aleksandra and I work on low-complexity regions of protein sequences. There are many ways of searching for them, and during the hackathon we determined that it would be useful to develop a tool comparing these methods and the results they generate. During the meeting, we developed the specifications for this software and agreed what it should include and how we want to approach the problem.

Aleksandra is responsible for the overall project. We have prepared a timetable, so that all participants know which tasks are their responsibility, what their deadlines are and what they need to hand over to the next person. And so we managed to create a full design during a single afternoon and morning.

And the participants went home to work on their own component of the project?
M.G.: That's right. The meeting lasted two and a half days. First, everyone delivered a 15-minute presentation on their own research, then we held the actual hackathon, and finally we worked on preparing better definitions of individual non-globular protein sequences. The following morning we honed down our ideas for future work. We were even able to discuss financing options for our ongoing collaboration – that's a lot to get out from two and a half days!

I have to admit I was hugely impressed, and the meeting resulted in a myriad of new ideas. We have a long list of tasks – we seem to have opened a bottomless vault, in a good sense. Every hackathon participant works on something slightly different, so we all complement one another really well.

Even in today's online world it's impossible to overestimate the value of meetings in person.
A.G.: That's right. There are many ways of keeping in touch, and we often work with Marcin remotely. We can go for months, even a year, without seeing each other, but sometimes it is essential to meet in person and talk. Our consortium is very interdisciplinary. Although we all work on low-complexity protein sequences, we have computer scientists, biologists, and everyone looks at the problem from a slightly different angle, which brings great results.

What happens next? How do you see the future of this project? Are you going to focus on the functions or applications of the proteins?
M.G.: This is a very complex question, because science – like everything else – is driven by fashions. The current fad is in RNA analysis, and researchers working in the field are awarded generous grants. Maybe the time will come for not just low-complexity sequences, but non-globular proteins in general. There are good reasons for this – they comprise around 15% of the total proteome, and they also remain largely unexplored.

So far we are a long way from any real applications – we are exploring completely unknown territory. There is little existing research we can work with and no concrete data we can use to verify our suppositions. It's simply because no one has been focused on this topic before.

Every hackathon participant works on something slightly different, so we complement one another really well.

You are exploring terra incognita.
M.G.: I think that our students are inspired by knowing that they are helping to open doors no one has opened before. This makes us all the more grateful to those who made the first inroads, including our partners from the Johannes Gutenberg University of Mainz who worked on homogeneous sequences, i.e. those comprising a single amino acid. There is also research into individual low-complexity fragments, or work concerning specific proteins. However, there is no global research similar to what Aleksandra and I are working on.
A.G.: There isn't yet, but we are working to make it happen!

INTERVIEW BY **DR. AGNIESZKA KLOCH**
PHOTOGRAPHY BY **JAKUB OSTAŁOWSKI**