SORTING OUT THE PIECES THAT MAKE US TICK

BY JASON TOGYER

BIOLOGY RELOADED

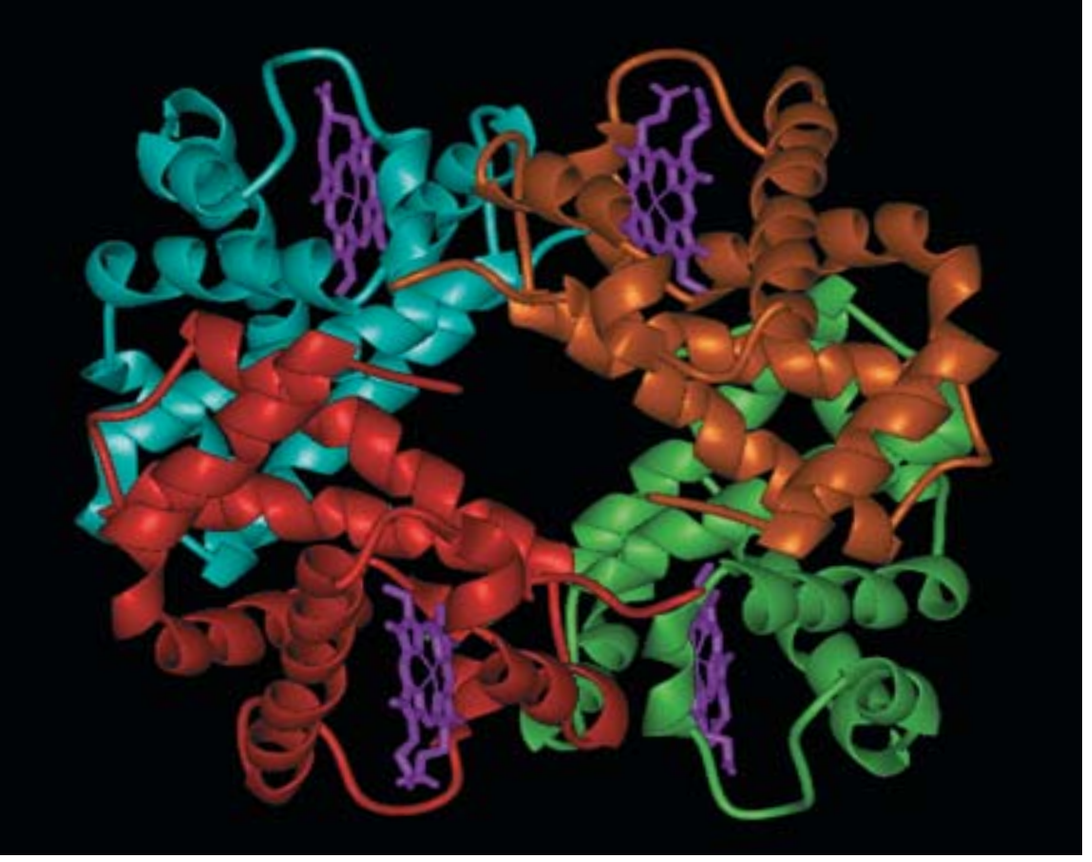
Imagine receiving a box containing 3 billion pairs of gears, springs, and levers. Inside are enough parts to build 24 working clocks and watches of various shapes and sizes. In the same box is a random selection of parts from windup toys and old typewriters. There are no blueprints to describe which parts fit together.

With enough experimentation, you might be able to construct several working machines. Some would keep time. The assemblages might even resemble the clocks from which the parts were taken. Still, even a skilled watchmaker would probably find the process frustrating and would explore many dead ends before creating something worthwhile.

This describes, roughly, the challenge presented to scientists by the human genome. The 24 clocks that can be assembled from the pile are the pairs of human chromosomes; the 3 billion pairs of parts are the nucleic acids—adenine, cytosine, guanine, and thymine (typically represented by the letters A, C, G, and T)—that come together to form DNA, the storehouse of information for encoding the tens of thousands of proteins responsible for most life functions.

Computational research team members at Pitt are ignoring what they hope are unnecessary biological data. They believe this approach will drastically accelerate our ability to cash in on the Human Genome Project, leading to new treatments for the enormous body of diseases and disorders caused by wayward proteins. FROM LEFT: **Hagai Meirovitch, Ivet Bahar, Takis Benos, and Dan Zuckerman.**

ILLUSTRATION | JOHN RITTER
PHOTOGRAPHY | C.E. MITCHELL



Last year, Bahar and postdoctoral fellows Dror Tobi and Chunyan Xu studied the movements of hemoglobin. They determined that relatively simple models of the protein could accurately and efficiently predict its movements. To do this they used "coarse-grained" structural information for the protein backbone, which is shown above in a ribbon diagram.

In April, 50 years after James Watson, Francis Crick, and Maurice Wilkins first described DNA's double-helix structure, researchers announced they'd completed sequencing the 3 billion pairs of nucleic acids in the human genome. With better than 99 percent accuracy, we now know the exact order of the building blocks of DNA.

That information alone has limited value, says Ivet Bahar, professor of molecular genetics and biochemistry in the University of Pittsburgh School of Medicine. When the sequencing process first began more than a decade ago, she notes, scientists thought they would eventually be able to go directly from the data they would glean to creation of new medical treatments.

"We soon realized that this gene-to-drug paradigm was not true," says Bahar, who heads the School of Medicine's recently created Center for Computational Biology and Bioinformatics. "It's not sufficient to know which genes exist or which are involved in a specific disease. We need to understand the machinery of the *proteins* encoded by these genes."

(Not to say mapping the human genome was unimportant. Pitt alumnus Lap-Chee Tsui, who in the late '80s helped discover the gene that causes cystic fibrosis, once told *The New York Times* that without a map of the genome, the work was like looking for a house in a city between New York and Los Angeles without a street address.)

Inside the cells of all living organisms, DNA interacts with probably 30,000 different kinds of proteins that carry out the biochemical reactions that keep cells alive, give them their unique characteristics, and allow them to reproduce. How a protein functions determines whether it can attach itself to DNA and other proteins, and what that attachment will look like. Some proteins don't attach, but serve as "signaling agents," triggering cascade reactions in other proteins in a cell.

When protein molecules, say, fold in the wrong place, cells go haywire. Many diseases are now believed to be caused by proteins that have the wrong structures. Some proteins trigger cascade reactions that cause cells to malfunction and multiply uncontrollably. We call that cancer.

The movements of protein molecules are vital to understanding genetic disorders and disease. If only we had drugs that could keep proteins from behaving in ways that cause malfunctions—then we could stop diseases before the first symptoms even appear, scientists believe.

"Each protein, to do its function, must undergo some motion at the molecular level," Bahar says. "It has to undergo some internal structural changes. These are like little molecular machines, and there are ways of increasing and decreasing the efficiency of these machines."

Given the three-dimensional nature of the interactions—all of these molecular machines whirring around—it makes sense to look at proteins not only by examining chemical reactions but mechanical movements as well. Yet conventional wisdom in research, until a few years ago, held that to understand the processes by which proteins function, we had to study activities solely at the atomic level.

That's like studying traffic patterns on a California freeway by analyzing the movements of every individual mechanical part in each individual car and truck—every piston and valve and bearing. The research would quickly be bogged down in details, some of which would be meaningful (the rotations per minute of the wheel bearings, for instance, which might give us the speed of the cars) and some of which would be worthless (the movements of the locks on the glove compartment doors).

What if we could look for patterns as groups of cars moved from lane to lane and from highway to off-ramp? Simultaneously, we might also look for anomalies—drivers who were speeding, ignoring traffic signs, and making left turns from the right lane.

To some extent, this describes the multilevel approach that Bahar has taken. She specializes in creating "coarse-grained" simulations of the ways that proteins interact. These computer models sacrifice detail on the atomic level in favor of more information about movements on the molecular level. Call it a "seeing the forest for the trees" approach.

"It was a brave start on her part in some ways, because she was taking a much different look at what people had done, and taking a look at much less detail than people had done before," says Robert Jernigan, former deputy chief of the experimental and computational biology lab at the National Cancer Institute in Bethesda, Md. He now directs the Laurence H. Baker Center for Bioinformatics and Biological Statistics at Iowa State University.

Jernigan says Bahar "ignores a lot of the details" that are not important at higher levels of protein functions.

Some details have to be ignored, because even the most powerful computers still get bogged down when trying to cal-

PROTEIN IMAGES: COURTESY THE CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

FORCEFUL INTERACTIONS

Half a century ago, a lab assistant at the University of Cambridge named Rosalind Franklin took the x-ray photographs of DNA strands that guided Francis Crick, Maurice Wilkins, and James Watson in deducing DNA's double-helix structure. Today, the method that Franklin used—called x-ray crystallography—is still the most efficient way to determine the 3-D structures of long molecules, including many proteins. Knowing the shapes of those molecules is vital to understanding how they work.

Yet x-ray crystallography has its limitations. As its name implies, crystallography requires the protein to be "crystallized"—"frozen" in an orderly fashion—and only a static picture of its structure can be obtained. However, in nature, proteins are dissolved in solution; and though they tend to prefer a well-defined structure, they are free to move. Pitt's Hagai Meirovitch demonstrates this by making two fists and placing them side-by-side, as if holding onto an imaginary bar.

"The protein chain has considerable freedom in space," he says, twisting and rotating his fists. "A large number of different 3-D structures can be formed by the flexible chain." That flexibility is essential to protein function.

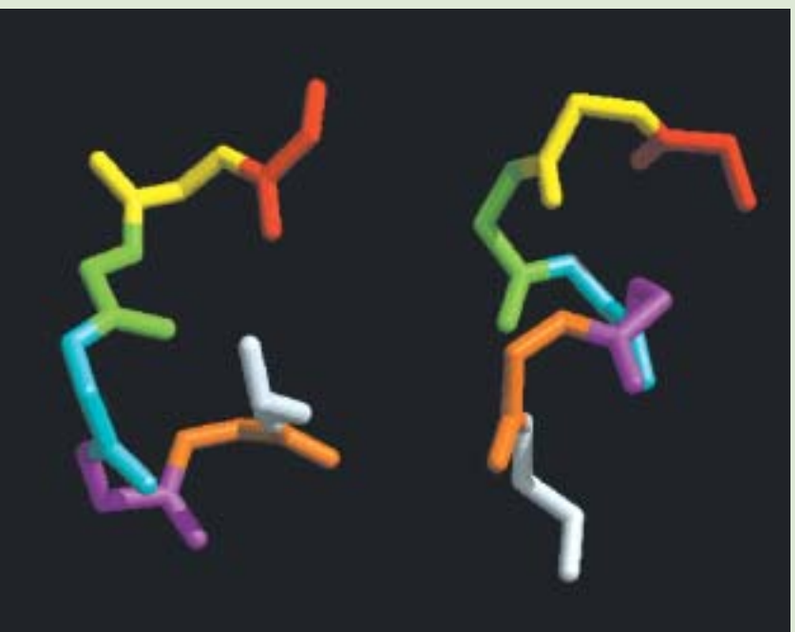
One thing researchers can do effectively is determine the sequence of amino acids in a protein. (For example, sequences of proteins that haven't even been determined experimentally are known from the genome project.) So, what if scientists could predict the structure and function of proteins based on their known sequences?

To do this, researchers would need a reliable mathematical description of the forces among all of the atoms involved—those of the protein and those in the surrounding water. They'd also need a way to simulate the movements of the molecules according to the laws of thermodynamics, taking into account tricky factors like entropy—the measure of order, or disorder, in a closed system.

Meirovitch has embraced this challenge. The professor of molecular genetics and biochemistry says there's much more work to do, yet his lab has created simplified models of protein-water interactions. In addition, he has developed computational methods for defining a protein's most stable structure. (A given protein's most stable structure also points scientists to its "active site," the location where certain chemical reactions occur most efficiently.)

His work has been helpful for studying segments of proteins called "surface loops." If the main protein body can be thought of as a few yards of bundled rope, imagine strands hanging off the sides on the outside. Those strands, or surface loops, are highly flexible and can act almost as feelers for the protein. (Sometimes they actually loop back to the molecule; sometimes they don't.) They play important roles in biological recognition processes such as antibody-antigen interactions.

Meirovitch says that tools like the ones his lab has developed will aid in the investigation of simple biological processes at the atomic level and the design of therapeutic drugs. "Stronger computers and improved computational techniques will enable [scientists] in the future to treat more complex problems," he notes. "Our mission is not just to develop methodologies but to apply them."



Meirovitch has developed simplified methods for determining stable protein and peptide structures. LEFT: The two most stable structures of a neural peptide known as deltorphin. The colors represent different amino acids.

culate the trajectories of tens of thousands of atoms, each changing direction thousands of times per second. "Computation is very, very time consuming," says Hagai Meirovitch, a professor of molecular genetics and biochemistry at Pitt, and one of three core investigators working with Bahar under the auspices of the Center for Computational Biology and Bioinformatics.

Because the human body is mostly water, reliable models of protein behavior and protein-DNA interactions must take into account the way the amino acids react to the water in which they're suspended, Meirovitch explains: "If you have 1,000 water molecules around the protein, each atom has an interaction with each of the 1,000 water molecules, plus the atoms inside the protein itself."

The volume of data that must then be processed means it's not unusual for computer simulations of molecules to run for *months* before useful results develop. And

Given the three-dimensional nature of the interactions, it makes sense to look at proteins not only by examining chemical reactions but mechanical movements as well.

the more data that is collected, the more complex those simulations become, says Kerstin Lindblad-Toh, a codirector of the genome sequencing and analysis program who led the mouse genome project at the Whitehead Institute's Center for Genome Research in Cambridge, Mass.

"We are clearly hitting issues with computing power," she says. Because computer technology and processing power continually improve, things will get better over time, Lindblad-Toh says, but science can't just wait for computers to evolve.

"It takes active thinking by the right people to make as simple a tool as possible so it takes the least amount of computing possible," Lindblad-Toh says. "We gain a lot by thinking about the most efficient way of doing things."

Computational efficiency is Bahar's goal, in the sense that she'd like scientists to have simpler, faster models of how DNA and proteins function. Understanding the chemical processes of the human body will allow "rational" design of drugs and vaccines, she says, rather than design by "trial and error." Since Bahar's arrival at Pitt in March 2001 from Bogazici University in Istanbul, she has recruited others, including

her fellow investigators at the center, to her line of thinking.

Bahar is unfailingly friendly, polite, and patient. With casual visitors, she seems somewhat reserved—colleagues who know her well, however, say that impression is misleading. Bahar's demeanor, they say, belies a mind that's constantly processing and an intellect that's intense and passionate. There's nothing mild about how Bahar approaches her work.

"Ivet is a real doer," says Ruth Nussinov, a professor of biochemistry in the Tel Aviv University School of Medicine and a principal investigator at the National Cancer Institute in Frederick, Md. She and Bahar met in 1994, when Bahar visited the NCI's Laboratory of Experimental and Computational Biology. "She is very energetic, very focused, and has always managed to accomplish a truly astonishing amount of work," says Nussinov. She points out the more than 150 journal articles that Bahar has published in the last 15 years.

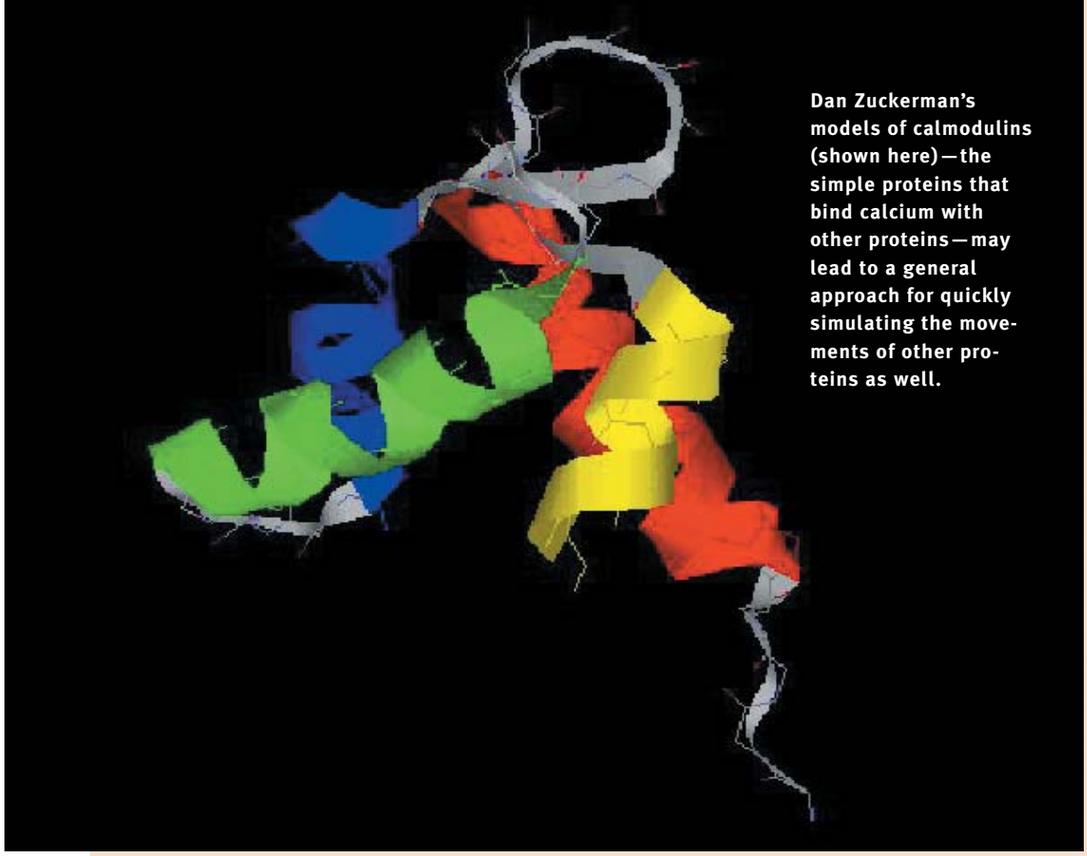
"I was always very impressed by her pace, and how quickly she [would] focus on a problem, [decide] how to go about it, do it, get results, and summarize them," Nussinov says. "If I remember right, every visit [by Bahar] to the [NCI] lab has resulted in at least two publications."

She has a way of putting her peers at ease and becoming deeply engaged in their work, says another NCI researcher, David Covell of the computational technologies laboratory in the screening technologies branch.

"She enjoys sitting down with people within her reach and very carefully going over all of the details of what they're doing," says Covell, who worked with Bahar on ways to model the flexibility of proteins. And although Bahar has advanced the study of coarse-grained models of protein behaviors, Covell calls Bahar's own behavior detail-oriented "in the extreme."

"Ivet goes to great lengths to ensure that you are treated hospitably," says Covell, who visited Bogazici University several years ago as a guest of Bahar and her husband. "The same attention to detail she puts into her science is the attention to detail she puts into her social interactions."

Bahar's research in recent years has touched on a wide variety of fields that all



Dan Zuckerman's models of calmodulins (shown here)—the simple proteins that bind calcium with other proteins—may lead to a general approach for quickly simulating the movements of other proteins as well.

PROTEIN IMAGE: COURTESY THE CENTER FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

FASTER, QUICKER, CHEAPER

For centuries, no one could figure out how a horse gallops; the movements happen too quickly for the human eye to discern. That's why so many early paintings show horses running with all four legs splayed out on the ground, like hobby horses. Then in 1872, California millionaire and racehorse owner Leland Stanford commissioned photographer Eadweard Muybridge to settle the debate once and for all. By arranging a series of remotely controlled cameras around a track, Muybridge was able to capture the intermediate movements of a horse's legs—and prove those early painters were incorrect. When a horse is running at full speed, all four hooves actually end up off the ground, but they never hit the ground splayed out hobby-horse-like.

Dan Zuckerman's work with protein molecules is on a substantially smaller scale than Muybridge's work with racehorses. But the principle is the same: To understand how bodies change from one state to another, we need to capture their intermediate stages of movement. The problem, Zuckerman notes, is that proteins can go through thousands of transitions every second. Modeling just one transition with conventional methods drains a tremendous amount of computing power.

"If studying these transitions requires hundreds of powerful computer processors, then the work is obviously limited to only a few scientists," Zuckerman says. "Given the number of important proteins that are being studied, the field can't advance very quickly."

On the other hand, if the model is able to run on a desktop computer, "then Joe Professor can do it," Zuckerman says. In other words, one would need the inspiration and know-how, but not access to large-scale computer processors.

This isn't just a hypothetical. In just two weeks, using the PC in his office, Zuckerman can generate more than 100 transitions in calmodulin, a fairly simple protein that binds calcium with a host of other proteins. Each of these transitions represents about one-tenth of a millisecond of calmodulin's ever-fluctuating motions.

Though calmodulin is "ubiquitous," says Zuckerman, he didn't set out to invent a method solely for analyzing it. He wants to create a general structural modeling approach for studying all sorts of proteins quickly. His solution to simplifying the simulations is to take a representative sample; instead of simulating the movement of every atom, for instance, Zuckerman's model might simulate every fifth or 10th atom.

"Maybe, if you throw away some of the data to get to the large-scale movements, you

are throwing away some interesting atoms," he says. But you can check the simulations against the available experimental information to see if the model's predictions are accurate. Certain very simple models can reproduce a surprising amount of data, says Zuckerman, who came to work with Ivet Bahar in September 2002 from a postdoctoral fellowship at Johns Hopkins University. "In the past, I had always worked on atomically detailed models," he says. Yet simulations of atomic models of proteins, which track the motions of tens of thousands of protein atoms, surrounded by thousands of water molecules, are stuck at extremely short time scales. A long series of simulations might have represented 10 nanoseconds in the life of a protein—not enough time to see anything, Zuckerman says: "Working here with Ivet has really opened my eyes to these simpler models."

That's like studying traffic patterns on a California freeway by analyzing the movements of every individual mechanical part in each individual car and truck—every piston and valve and bearing.

SIGNALS AMID THE NOISE

In the early 1990s, when the race to decode the human genome began, it cost about $10 to identify a single base pair. Back then, a technician could manually scan about 10,000 base pairs per day. At those rates, it would have taken a team of 20 technicians about 40 years and $30 billion to sequence the human genome.

Improved technology has lowered the cost to 5 cents per base pair; the leaders of the Human Genome Project estimate the final cost of sequencing the human genome at the bargain rate of $2.6 billion. And today's automated laboratory equipment can scan 10,000 base pairs per second. Not only are the new processes cheaper, they allow greater accuracy, because technicians can check and double-check sections of the genome.

To fully exploit this wealth of data, we need more efficient algorithms, says Pitt's Takis Benos. Many scientists are exploring how protein-coding and noncoding genes function; Benos is interested in how gene regulation is fine tuned. A change in a single base pair, for instance, could result in a gene being misexpressed, and that can translate to a serious disease. His laboratory is developing algorithms to scan the human genome and detect short but important sequences driving gene expression. They compare the human sequences with other species, like mice. "We expect that because the unimportant DNA changes quickly, this comparison will reveal the important parts," he says.

But finding this information is not easy. As Benos explains it, it's like sitting on the back porch late at night, tuning a shortwave radio. We patiently turn the dial, passing up squeals and rushes of static, until faint music or voices can be heard. "We're looking for faint signals amid the noise," Benos says. In genomes, he adds, "the signals that are important are relatively short, say six to 12 bases, and they are surrounded by long strings of genetic noise"—or base pairs that don't encode important information. That noise can generate a lot of false positives, notes Benos.

To cut down on these false positives, Benos is applying statistical methods to make a kind of spot check, taking a representative sample of genetic data. The formulas with which he is working dig through data from two or more genomes of different species, looking for similarities and matching patterns, then ranking the results to see if they might be important. In the case of our shortwave radio example, it would be like having two friends in neighboring towns tuning their radios randomly, then telling you about frequency ranges where they found what seem to be good programs. You would have a higher probability of finding something you liked this way than you would have without your friends' help.

In tuning out the genetic noise, another challenge for Benos: From the beginning of a gene in a genome, how far out does he keep looking for promising signals? "In a simple organism like yeast, 500 or 1,000 bases are usually sufficient," he says. "In a mouse or human, how far should we go looking? Five thousand? Ten thousand?"

fall under the general label of "computational biology." Her studies—and similar work being done elsewhere—could one day lead to better treatments for the enormous variety of disorders caused by protein misfolding and aggregations as well as for diseases caused by protein signaling and regulation mishaps (which are instrumental to the development of cancer). Yet Bahar's background isn't in biology or medicine. A chemical engineer by training, Bahar began her career studying polymers but found life sciences "more interesting than producing high technology chemicals for industry."

The engineer recognized that many of the methods, tools, and fundamental concepts from the world of synthetics could be applied to biological molecules.

"Her background is a great asset," says Nussinov, noting that Bahar's engineering training makes her well-suited for carrying out complicated mathematical calculations on biological molecules.

"In this respect, it's an infinitely better background than biology—and I know that for a fact, since my background is in biology," Nussinov says.

The principal researchers in the Center for Computational Biology and Bioinformatics share Bahar's varied interests and, like her, aren't traditional biologists. Meirovitch and Daniel Zuckerman, an assistant professor of environmental and occupational health in Pitt's Graduate School of Public Health (GSPH), examine problems of structural biology—determining the shapes of proteins and studying their motions and potential interactions. Zuckerman is a physicist; Meirovitch is a physical chemist. Takis Benos, an assistant professor of human genetics in GSPH, mines data—digging the most important facts out of mounds of unsorted information. His degrees are in mathematics as well as molecular biology.

The mix of disciplines Bahar has selected for her team is well-chosen, says Whitehead's Lindblad-Toh:

"We're going to need many different fields coming together to pull out this knowledge.

"We need physicians and biologists to ask, 'What is important to medicine for us to look at?' But we need computational biology to determine how." ■