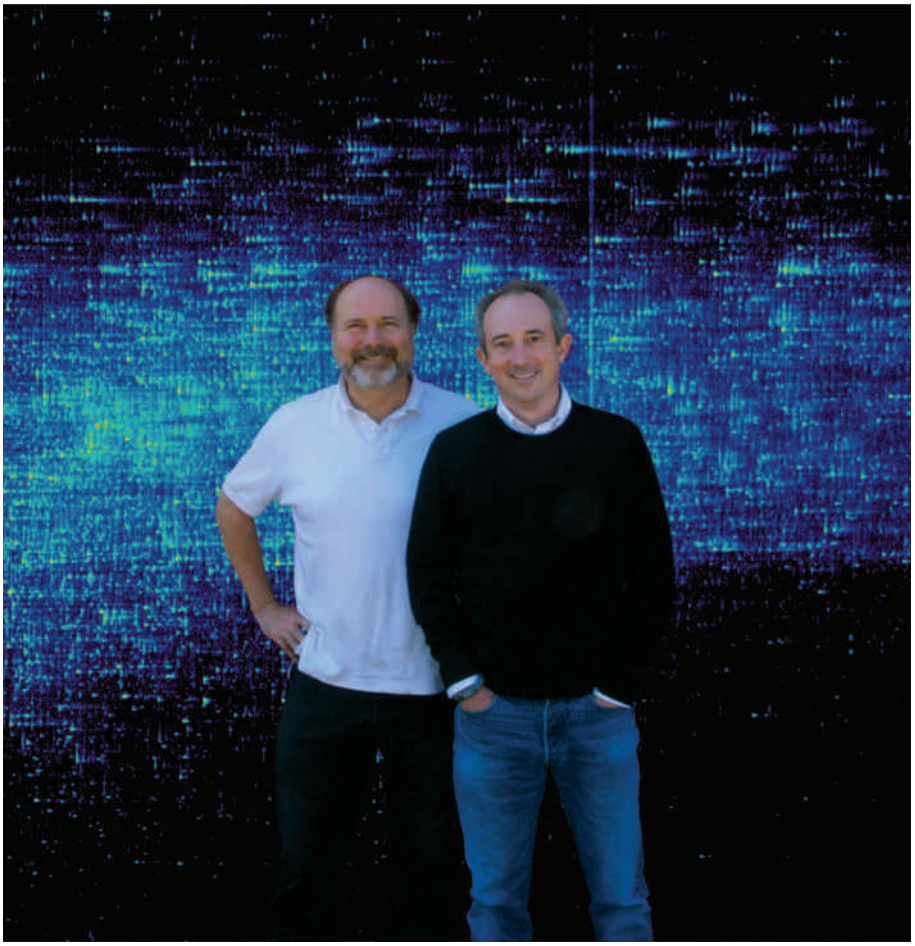


APPLIED PROTEOMICS

Danny Hillis (left) and Davis Agus are hoping to use protein signatures to diagnose and improve the treatment of cancer.

MEGADATA

The odd couple

An unlikely duo is trying to make sense of the avalanche of data that confronts cancer scientists, pointing the way towards a new era of research.

BY ERIK VANCE

In March 2004, oncologist David Agus was leaving his hospital in West Hollywood, California, at the end of a long day. On a whim, he cast his eye over a few magazines in the gift shop as he walked past. When he got to *Fortune* magazine, he stopped dead in his tracks. The cover picture was of an ominous, translucent cancer cell overlaid by the words: "Why We're Losing the War on Cancer". The story took the entire cancer research community to task and shook Agus to the core. It claimed that despite 35 years of work and more than US$200 billion spent on US research alone, the mortality rate for cancer patients across the world had barely budged. These were damning words to a man who had spent his entire career trying to cure patients of the disease and finding ways to stop it.

"It was very powerful," says Agus, now an oncology researcher and physician at the University of Southern California in Los Angeles. "At first I was crestfallen. Then it was almost a call to arms."

The *Fortune* article laid out a series of penetrating critiques of cancer research and called for a wholesale restructuring of President Richard Nixon's 'war on cancer', which had created the National Cancer Institute (NCI) and has guided the US approach to cancer research ever since. The notion that cancer research was somehow misguided would stay with Agus, gnawing at him.

He eventually found an outlet for his frustration through proteomics — the study of the body's vast ecosystem of proteins known as the proteome. And he found a partner from outside the medical world who was adept at wading through the unthinkably vast reams of data coming from proteomic investigations. Together, the pair set about creating a computer model that could contrast a healthy proteome with those of identified cancer patients to find biomarkers for the disease. Their goal: to diagnose a cancer patient before his tumour grows. After eight years, Agus feels he has finally answered the charges levelled at researchers in the *Fortune* article — but it will be a few more years at least before he'll know whether those answers can change the fate of his patients.

To Agus, the most critical, and personal, of the *Fortune*'s accusations was that the cancer community had become reductionist, too focused on the minutiae of just a few genes and pathways. It was a revelation of sorts. He had published many papers along these lines, but was finding cancer just too complex, with too many changes to the body's systems occurring outside the genetic code. Finding ways to block these pathways had led to several innovative treatments, but blocking one route often forces the tumour to grow a different way.

The article strengthened doubts Agus was already beginning to have. "These complex emergent systems are impossible to understand," he says. "Our level of understanding is just so cursory that we have to start to look for what they call, in physics, coarse-grained elements."

For a biologist, a complex emergent system is seemingly disorganized; its constituent parts tell you little about how it functions as a whole. Coarse-grained, in this context, means shifting the focus away from those individual parts to look at multiple scales: for instance, at the level of the cell, the tumour, and the entire system. Viewing data this way is like stepping back from a pointillist painting to take in the whole scene, rather than focusing on individual dots of paint. Similarly, stepping back from a cell makes it obvious that it's not the genes that do the heavy lifting, but the proteins they encode, Agus says. Cancer may be less about genetic mistakes than about malfunctioning proteins. But the proteome is far more varied and complex than the genome, so analysing it is a daunting proposition.

THE PROTEIN PROBLEM

"Twenty years ago, 'proteins' was a dirty word," says John Blume, a molecular biologist and chief scientific officer at Applied Proteomics in San Diego, California, the company that Agus co-founded. "It was a very black art. You poured gels, you did this sort of incantation over them, and they turned blue or they turned

black or whatever — it was like trying to get things to materialize in the witches' cauldron, whereas DNA was very digital."

It is hard to overstate how much more complicated the proteome is than the genome. The genome is coded by four nucleic acids that are always the same, making analysis of the genome a problem well suited to computer analysis. Proteins, on the other hand, are constantly changing. The same proteins with the same chemical formula will have different functions depending on their shape.

But over the past few decades, advances in mass spectroscopy have allowed researchers to identify chemical components more accurately. The ability to process massive amounts of data has also improved. This combination has convinced many scientists that the time has come to pull the proteome out of the witch's cauldron and into the daylight.

Agus wanted to create a whole-body picture of the proteome at any given time using just a sample of urine or blood. From that picture, a doctor could theoretically identify a group of proteins indicating a nascent tumour. But the system is so complex that finding such a tiny anomaly would be like looking at a global climate model to determine whether a rain shower is coming to your town.

So Agus needed help. But what came next sounds more like the beginning of a bad joke than a fruitful collaboration. "I was in my lab and [former US vice-president] Al Gore came in," he recalls. "I was showing him what we are doing — proteomics and all this. And he said: 'Agus, this really would benefit from having an engineer's way of thinking attached to it'." According to Gore, there was too much pure science and not enough problem-solving going on, Agus says. Gore thought he knew where to turn for help. He suggested that Agus call Danny Hillis, a pioneer of parallel-processing computing and former vice-president of Walt Disney Imagineering.

But Hillis was not keen to talk. "Being in the business I am, I get calls from a lot of different people," he says. "Calling [Agus] back wasn't a high priority because I knew how difficult the problem was." Then one day Hillis came into the office to find messages from Gore, venture capitalist and presidential adviser John Doerr, and entrepreneur and investor Bill Berkman, all of whom said, basically: "Call this guy back."

Hillis possesses an ideal combination of skills to help Agus: he is both a wizard with complex information systems and an exceptional businessman and entrepreneur. In the early 1980s, while working on artificial intelligence, he created one of the first parallel supercomputers, which grew into a $65 million company called Thinking Machines. At Disney, he created business strategies, theme-park rides and giant robot dinosaurs. His current company, Applied Minds, based in Glendale, California, works with large organizations

MARRINUCCI, D. ET AL. PHYS. BIOL. 9, 016003 (2012)

BIG DATA, BIG BUSINESS

Hillis and Agus are just two members of a growing field of protein hunters.

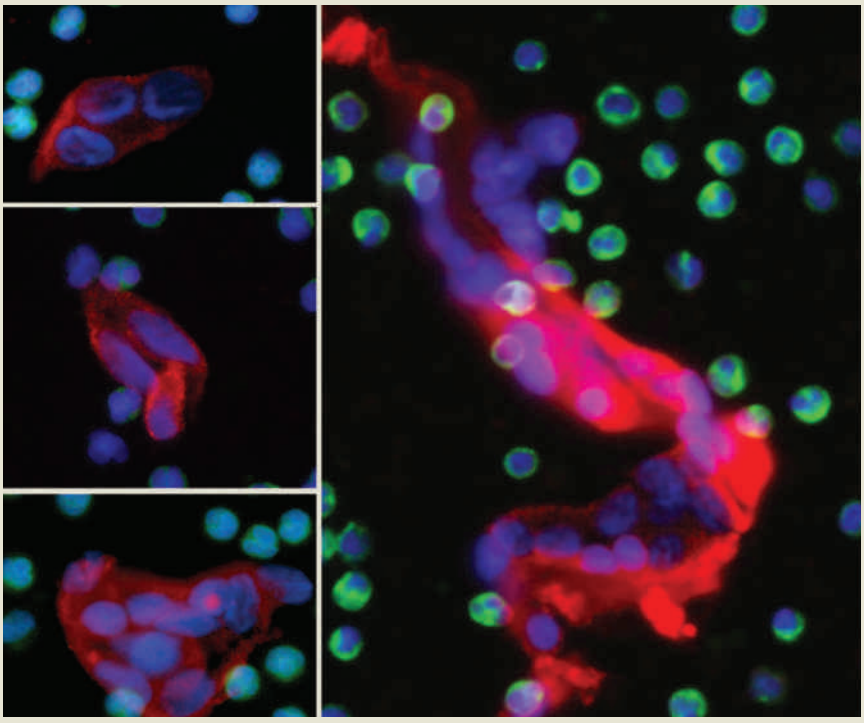
The shift towards using high volumes of protein data for medicine is not unique to Applied Proteomics. Several companies and academic labs are looking carefully at diagnostics along similar lines.

Nodality in South San Francisco, California, was founded on the premise that medicine is failing to accurately categorize patients according to which therapies might work best. The company uses a patented technique, created at Stanford University, that evaluates how cells communicate with each other. Nodality's business is based on building diagnostics, but most of its team has a background in therapeutics, so it focuses on how to match patients with specific treatments. Its first product is a test to match acute myeloid leukaemia patients with appropriate therapies.

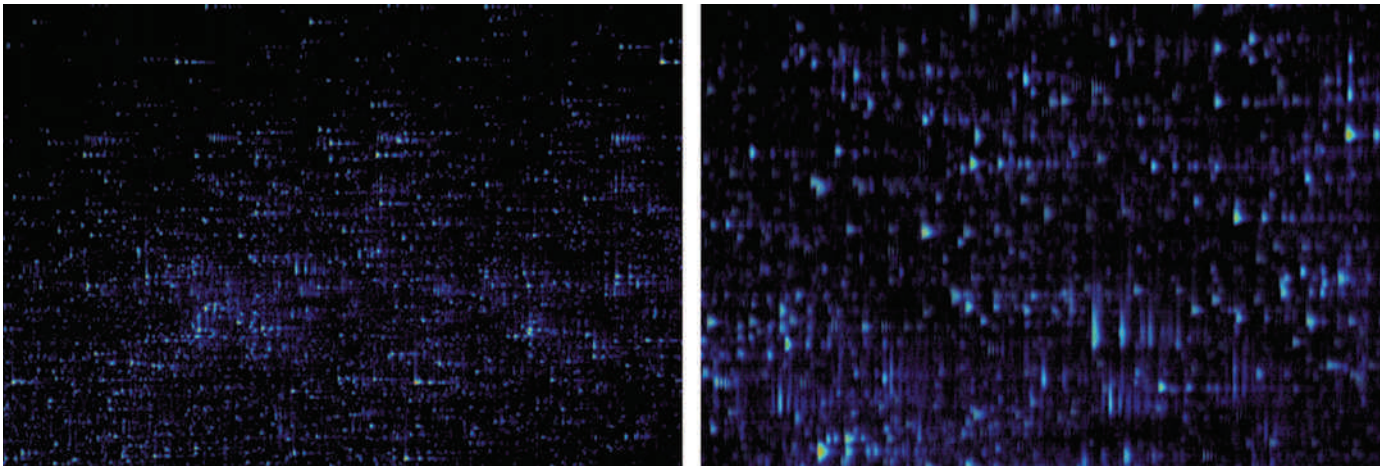
Integrated Diagnostics in Seattle, Washington, is developing blood-based tests for the early identification of cancer, diabetes and Alzheimer's disease. The company's innovation is synthetic antibodies dubbed 'protein-catalysed capture agents', which bind to specific molecules that are markers of various diseases. These synthetic antibodies weigh much less than an equivalent antibody, and are more stable and less expensive than other antibodies currently used in diagnostics.

D. E. Shaw Research in New York is a computational biochemistry company set up by David Shaw, a computer scientist who made his fortune by founding a lucrative hedge fund. He swapped Wall Street for the laboratory, and in 2008 built one of the world's most powerful computers to analyse the dizzying array of ways that proteins fold. His research group also created open-source software for analysing simulations of molecular dynamics.

Epic Sciences in San Diego, California, is a cancer diagnostics company that is investigating protein and genomic biomarkers for the early identification of prostate, lung, breast, pancreatic and ovarian cancers. It is participating in 12 clinical trials in the United States, Europe and Asia as part of 40 projects looking at 15 protein or gene tumour markers, says chief executive David Nelson. ***E.V.***



Blood samples from cancer patients contain normal blood cells (green and blue) and clusters of circulating tumour cells (red and blue). Epic Sciences is working to identify these circulating cancer cells and characterize them using protein or genomic markers.



Visualizations of proteins help analyze hundreds of thousands of features simultaneously, revealing both the big picture and smallest details to find disease-specific patterns.

effectively as an independent R&D department. Hillis calls it a "secret laboratory", where engineers design everything from a voice scrambler, to shield conversations from the people sitting next to you, to vehicles for use on other planets.

Harnessing the proteome would be a monumental challenge, however, and Hillis knew that. Not only are there two million proteins with almost limitless configurations, but their concentrations in the blood vary wildly, even from one hour to the next. One protein might be 10,000 times less common than another and yet hold the key to diagnosing and understanding the manifestations of cancer. "I had looked at this before and decided it was just technically too hard," Hillis says. But talking to Agus convinced him to try. "I got to understand how he was thinking about cancer and came to realize just how important this information was," he says.

BUGS IN THE CODE

For two years, Agus and Hillis worked under the auspices of Applied Minds on the unwieldy proteome problem. Hillis used an engineering approach: DNA is like code, and cancer is a bug in the code. To fix the bug, you must not only understand the code, but also understand how it interacts with the computer. In biological terms, the researchers believed that the way proteins communicate between cells is more important than how they are coded for by DNA. It's hardly a novel idea; scientists have been searching for more than a decade for cancer proteins travelling through the blood that might act as biomarkers. But according to Hillis, protein hunting — even in healthy individuals — was almost impossible. Two labs could test the same blood sample and could get very different results.

So Agus and Hillis set about building a robot assembly line that made procedures previously done on individual lab benches precisely replicable. Hillis compares it to semiconductors, which only became economically practical when they could be made in bulk with precision.

At the end of those two years, the pair founded Applied Proteomics to create a protein diagnostic that reveals not just where a cancer is, but how it interacts with the body. Statistically speaking, the task is akin to analysing climate or earthquake data, as it involves sifting through massive amounts of noise to find a signal. Perhaps a closer comparison would be making sense of a functional magnetic resonance imaging (fMRI) brain scan. Unlike fMRI, however, which measures blood oxygen levels in the brain, proteins are not limited to the skull but found throughout the body. Given the sheer number of proteins, the complexity of each one, and their widely varying concentrations, this analysis is a massive task.

If the model works, Agus and Hillis may be able to move beyond diagnostics. If cancer proteins can be separated out from the noise of the human proteome, doctors could tailor therapies to an individual's chemistry and target therapies for each individual cancer. With a full picture of the proteome, drugs that were previously considered failures — those that cured only a small percentage of patients — could be used in a suite of treatments tailored to individual chemistry. But this remains a distant prospect; for now, researchers must still contend with massive data loads.

"We are going to get overwhelmed by this enormous amount of genomic information," says Robert Austin, a Princeton University physicist who studies the way tumours behave and evolve in the context of their environment. "It's sort of like high-energy physics before the quark model came about," he says. "We lack a 'theory of cancer' right now."

MESSY DATA

Other scientists have previously tried to find patterns in the human proteome but with little success. The NCI has launched several efforts to catalogue the proteome, and the Obama administration is calling for a 'proteome project' analogous to the Human Genome Project, but progress has been disappointingly slow.

Applied Proteomics is the latest in a series of outsiders trying to shake up the musty halls of medicine (see 'Big data, big business'). The Human Genome Project also integrated physicists who were comfortable with vast amounts of data. John Quackenbush, for example, was drafted into the genome project from the high-energy physics lab Fermilab, where he worked with terabytes of data in the 1980s. But he found that his controlled physics experiments could not be more different from the messy, chaotic world of the human body.

"Bringing somebody in with a new perspective and a new way to think about a problem can be very useful," Quackenbush says. "On the other hand, I can tell you from my experience that sometimes coming in like that you can be a little naive in thinking that having a lot of data will suddenly solve all of your problems. Big data is not a panacea."

"It's sort of like high-energy physics before the quark model came about. We lack a 'theory of cancer' right now."

Neither Agus nor Hillis think this coarse-grained approach to diagnosing and treating cancer will be easy or guarantee new therapies. Yet when they talk about their work, they slip into broad, sweeping statements about the future of medicine.

"We are obligated to do things differently because the current method hasn't worked," Agus says. "It's not like we are making advances every year and things are hunky dory." And he believes that progress depends on interdisciplinary collaboration. "There is this notion that this is the century of biology — well, that's poppycock," he says. "This is the century of the convergence of the sciences." ■

Erik Vance *is a freelance science journalist based in Mexico City.*