NATURALISM IN THE COMPUTER AGE

A field guide to computational biology

CHARLES DARWIN WOULD never have predicted that his disciplinary decendents – biologists driven by a passion for exploration and observation of the natural world – would do their most productive work in an office. But a rapidly growing army of modern day naturalists focuses on understanding the complex details of the biological world through an exploration instrument highly divergent from Darwin's *Beagle* – the desktop computer.

Computers have revolutionized the way we meet, carry out business, tell jokes, share photographs, and pay our bills. It is no surprise that they have also radically changed the landscape of scientific research. Though not all experimentalists have packed up their benches to make room for rows of processors, a considerable number of scientists are relying on computers as research equipment. **Burkhard Rost**, a bioinformaticist at Columbia University, predicts that computational tools are so powerful that "no experimental lab will live without [them] by the end of this decade."

The overarching field encompassing the application of computers to answer biological problems has been dubbed *computational biology*, although, as with all emerging disciplines, there is considerable murkiness surrounding the label. You would be hard pressed to find a 21st century scientific laboratory that does not rely heavily on computers. But searching PubMed online for the latest research paper or running a microscope from a desktop machine does not a computational biologist make. Computational biology, along with its close relative, bioinformatics, is a science whose practitioners use computerized theoretical models as their primary research tool.

The Computational Biology and Bioinformatics Discussion Group, one of a diverse set of groups that participates in the Academy's Frontiers of Science program, never fails to intrigue its members, not least because of its unpredictability. The questions computational biologists seek to answer are as broad as the problems that constitute all of biology. The common link is not the nature of the questions, but the approach to answering them. Still, a handful of problems have emerged as the centerpieces of computational biology and it is worth noting that each deals with a complex system and an enormous amount of data. For a brief tour of the state-of-the-science and a small taste of the past year's talks, keep reading.

The questions computational biologists seek to answer are as broad as the problems that constitute all of biology. The common link is not the nature of the questions, but the approach to answering them.

MINING THE TINIEST GEMS

When asked to name the key advantages of using computational approaches to answer biological questions, Cold Spring Harbor Laboratory molecular biologist **Michael Zhang** does not hesitate: "Speed, economy, and necessity." Such an incisive response resonates with researchers working on one of today's biggest biological problems: understanding human health and treating disease. The past few years have brought great leaps forward in this endeavor. Most strikingly, the completion of the human genome project (which itself relied on serious computational prowess) has brought scientists a vast amount of information that holds the very secrets of humanity. Bioinformaticists – who specialize in the mathematical analysis of large data sets – are working feverishly to mine this data to discover the inestimable gems it holds.

Although we have collected the information we need to understand life, it is encoded in layers of complexity – a triumvirate of sequences (DNA, RNA, and protein) stores the instructions for the molecules that regulate life processes. One method of mining this information is to create sequence libraries – raw DNA from cancerous tumors, RNA coding regions, amino acid recipes for proteins – and enlisting computers to chug through the vast amounts of data they hold to find patterns, repetitions, or other interesting anomalies.

Put simply, sequence libraries are enormous databases filled with the codes that underlie nature. Researchers who employ brute force approaches to crack them will likely fall short of a breakthrough. Analyzing this vast data requires creative, intelligent analysis, and, more often than not, the development of clever mathematical algorithms. **Laxmi Parida**, of the IBM T.J. Watson Research Center, developed a sophisticated mathematical technique for identifying important patterns in protein sequences. "Once biology went molecular it paved the way for computational approaches," she noted,

alluding to the sequences of biological molecules that can easily be converted into digital strings of 1s and 0s for computers to read. Her statistical method seeks to demystify the large data set of protein sequences, which determines the ultimate structures and functions of the molecules responsible for the diversity of life.

Life's diversity is what inspires MIT's **Christopher Burge**, who uses messenger RNA libraries to understand the elusive rules of gene expression. To shed light on how a small set of genes is carefully regulated to achieve a wide range of purposes, Burge's lab set out to create a library of exonic splicing silencers – short strings of RNA that turn off a particular gene. Such a library provides scientists with a reference for understanding the vitally important RNA splicing code, which edits and translates DNA instructions into protein products. Powerful biological insight, like the peek of nature that Burge is privileged to view, has propelled scores of mathematicians and computer scientists to move into the world of biology, creating an explosion in the field of bioinformatics in recent years.

MODEL MOLECULE

One of the classical rules of scientific research is that if you can replicate a phenomenon, then you truly understand it. Car mechanics well know that there is no better way to grasp the inner workings of an engine than to take one apart and then rebuild it. The most important machines in the world are proteins and nucleic acids, the molecules that direct life. Even though they are unimaginably small, they are orders of magnitude more intricate than any car engine. The size and complexity of objects worth understanding in biology render most of them impossible to comprehend through experimental tools.

This is where computers come in handy. With a few strokes on the keyboard – after some sophisticated programming – computational biologists can create models of interesting systems and simulate their activity. A movie of a protein folding into its native state can provide an incredible degree of insight into the function of a real protein, which can lead to the design of therapeutic drugs. Comments **Tamar Schlick**, an applied mathematician who directs the computational biology doctoral program at New York University: "Modeling and simulation allow us to experiment with situations that are unfeasible, too expensive, or too dangerous to explore in the wet laboratory."

Molecular modeling is one of the cornerstones of computational biology, evidenced by the number of scientists who speak on this topic to the discussion group. **Wilma Olson** of Rutgers University uses computational modeling to understand the physical and chemical properties of the nucleosome, a deformed region of the chromosomal scaffold. A theoretical model of the electrostatic interactions in chromosomes can be worked out with a pencil and paper: DNA's phosphate backbone is negatively charged and the amino acid side groups of histones (the nucleosome's protein components) are positive. Everybody knows that opposites attract, but try scaling up to take into account hundreds of thousands of atomic interactions and you will quickly run out of paper and patience. Utilizing computing power to get to the bottom of the nucleosome's architecture, Olson discovered that charge neutrality was not, as expected, the prevailing force dominating the deformation. In fact, the actual sequence of DNA seems to be encoded in order to best twist

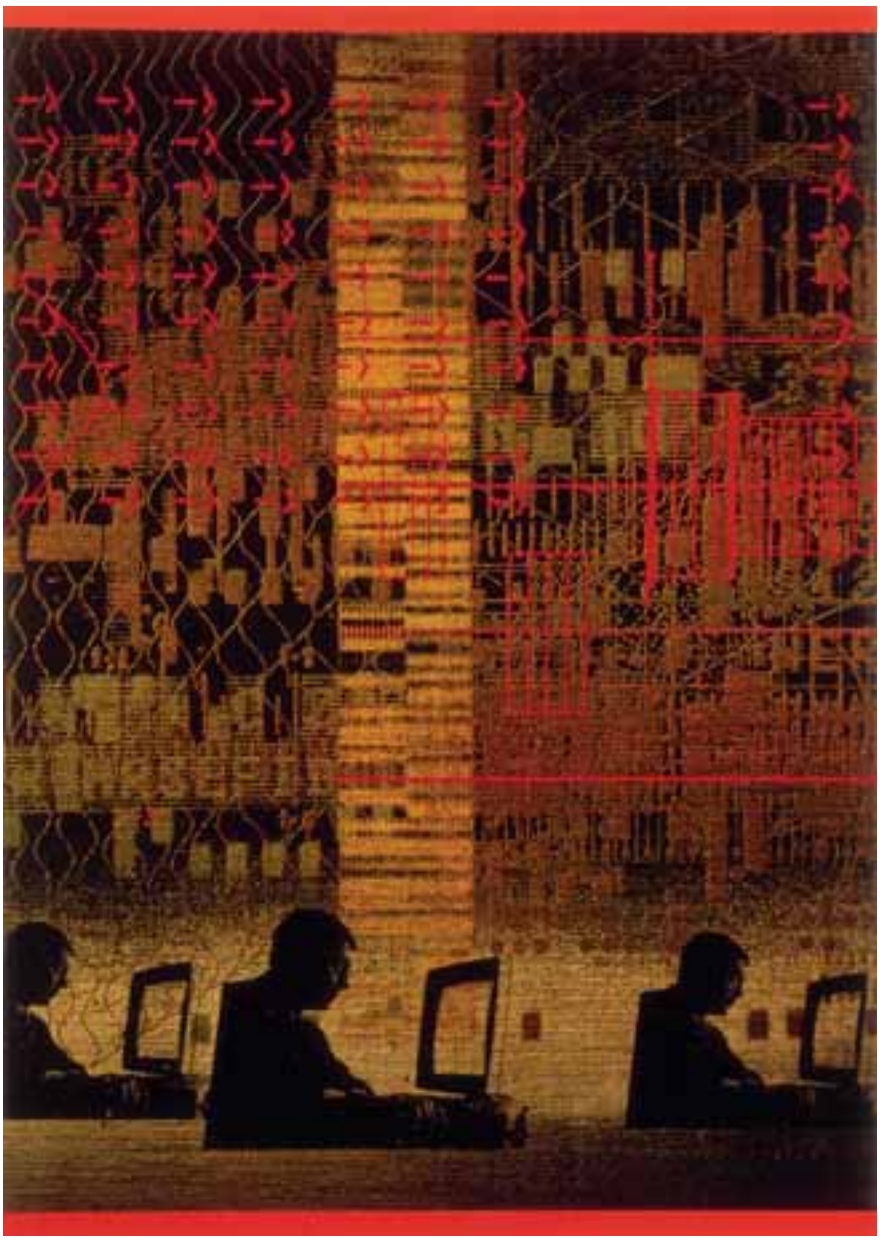
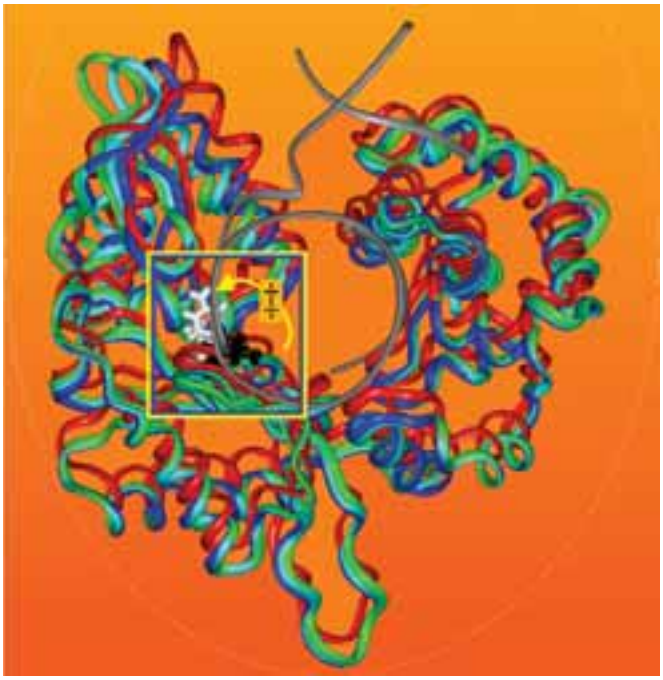


ILLUSTRATION: FRED OTNES – IMAGES.COM



Molecular modeling provides a view of molecules that are too tiny to see, even with a microscope. This protein-DNA complex (DNA polymerase beta with primer/template DNA) was simulated by Tamar Schlick's computational biology group at NYU (by Ravi Radhakrishnan and Karunesh Arora) to understand at the atomic level the mechanisms that regulate fidelity, or the faithful duplication and repair of DNA.

into the stable nucleosome structure.

In his presentation at the Academy, Princeton University's **Ned Wingreen** described a minimalist model he created to mimic protein folding. Where Olson focused on charge and the electrostatic forces inherent to nature's building blocks, Wingreen took into account only the hydrophobic effect, looking at how different protein residues interact with water. His goal was to determine the designability of protein structures, and his substantial approximations were justified when he compared his theoretical results to the structures of known proteins.

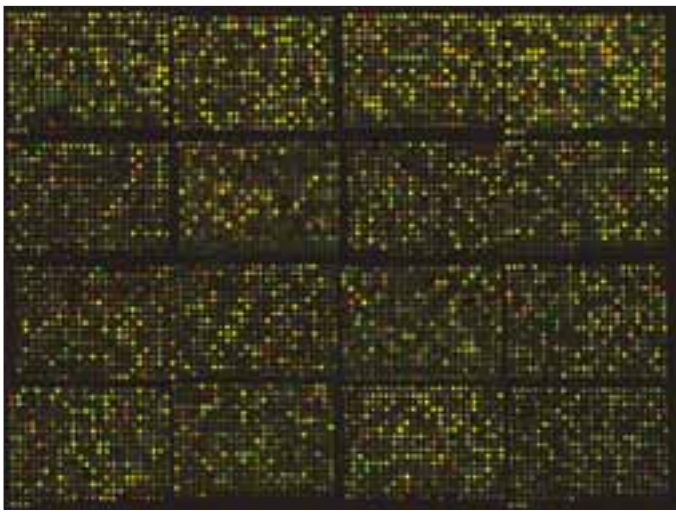
Wingreen's reductionism points to one of the key challenges in modeling: the three-way trade-off among the competing computational currencies of time, system size, and accuracy. Stating reasonable assumptions about the natural world that are accurate enough for your system of interest without being too computationally expensive is a central source of tension in computational biology. Just as Newton's simple laws of motion can pinpoint the arrival time of a transatlantic flight by considering the speed, distance, and frictional forces such as air resistance, quantum mechanics provides us with a set of laws that can accurately predict the movement of molecules in the natural world. Unfortunately, the math that describes the interactions between particles is so intensive that it is not practical to simulate a biological system in all its glorious detail, unless you have several lifetimes to spare. Assumptions must be made, and determining what approximations are appropriate for answering a specific question about a particular system is sometimes more of an art than a science.

Computational modeling is artistic in other ways, as anyone who has seen a molecular movie will affirm. Such colorful renderings provide a visual image of chemical events under experimental scrutiny as well as a mathematical model to explain confounding results. **Steven Schwartz**, a theoretical chemist at the Albert Einstein College of Medicine, points out that computation does not only bolster experiment, but can also surpass discoveries at the bench. "There are some questions that just cannot be answered with experiment – the enzyme mechanism question is one where computation has brought basic new ideas to the field."

Likewise, predicting protein structures, a task that often daunts experimentalists, presents a fertile opportunity for computational biologists interested in fighting diseases like cancer. **Ronald Dunbrack** of the Fox Chase Cancer Center tries to predict the unknown structures of proteins through a technique known as comparative modeling. Dunbrack pilfers theoretical methods from physics, computer science, mathematics, robotics, and quantum field theory, and cleverly applies them to better understand the molecular basis of cancer.

PLAYING BATTLESHIP IN THE GENE AGE

One of the unparalleled successes of computational biology has been the development of the gene chip or microarray. Modeled on a silicon chip, this marvelous device separates individual genes or gene regions into the thousands of spots on a tiny grid that looks something like a Battleship gameboard. As researchers deposit DNA or RNA onto the chip, components line up specifically at the holes where similar DNA or RNA is located. This enables scientists to assess all the genome's DNA or all the RNA transcripts simultaneously, providing a genome-wide picture of gene expression. A microarray view of an individual's DNA can help decode which genes give rise to cancer or other diseases. Bioinformatics laboratories like that of **Olga Troyanskaya** at Princeton University, use microarrays to identify mutations in tumor DNA. Specifically, Troyanskaya searches for copy-number changes – nat-



This microarray contains 6,000 spots, each of which holds a small gene region of the parasite *Plasmodium falciparum*, the cause of the most deadly form of human malaria. Joseph DeRisi at UCSF analyzes the lit regions on the chip to understand how genes are regulated.

urally arising amplifications or deletions of genes – which may cause disease.

Microarray data is analyzed through highly sophisticated mathematical algorithms, and many bioinformaticists are mathematicians and computer scientists who find their way into the messy world of biology. Columbia University's **Harmen Bussemaker**, a one-time theoretical physicist, studies gene expression through RNA microarrays. At the Academy a graduate student in his lab, **Barrett Foat**, highlighted the group's reverse engineering approach to identifying the noncoding RNA sequences that control genes, turning them on or off.

Since its development the mid-'90s, the microarray has become the workhorse of bioinformatics research. It has also set a standard for data-driven computational biology, forcing scientists in more traditional fields to perk up and discover the power of computation. **Chris Wiggins**, a theoretical physicist-turned-computational biologist at Columbia, pointed to the transformative power that microarray technology has exhibited on the field: "I think microarrays have shown how well biology can be revealed through data, and have convinced biologists that they benefit from talking to numerically minded people. Now it's time to take a 'high throughput' approach to the rest of biology – microscopy and other technologies should be amenable to these same data-driven approaches."

COMPUTING COST AND BENEFIT

Microarrays, simulations, and sequence libraries have irreversibly altered biological research with stunning findings about nature and health. Often the most exciting discoveries in computational biology are not concrete results, but novel methods and algorithms whose power lies in their versatility to be applied to a variety of problems. As a result, newcomers to the field may find it dauntingly technical, requiring a higher level of mathematical literacy than many other biological disciplines.

NYU's Tamar Schlick admits, "It's much easier to be a mathematician, chemist, or computer scientist, than a computational biologist who must employ all these disciplines in biological research." Yet the field is exciting enough to attract individuals with backgrounds in each of these distinct disciplines willing to learn the languages and cultures of all of the other research specialties. Chris Wiggins highlights the collaborative interdisciplinary appeal of computational biology, "Because we are all coming from different backgrounds, all of us not only speak multiple [scientific] languages, but we know how to communicate in a language that doesn't presuppose that somebody else already knows ... [a] particular piece of technical jargon." At the Academy, biologists from more traditional groups have been opening dialogs with computationalists, knowing that computers will no doubt foster the future of discovery.

The emergence of computational biology as a powerful interdisciplinary research area has vastly altered the biological sciences, opening new routes for the exploration of the natural world. As the field continues to evolve and more researchers learn computational techniques, scientists will be reminded that at the core, computational biologists are just naturalists for the 21st century.

–Kiryn Haslinger

2005 Highlights from the NYAS Computational Biology & Bioinformatics Discussion Group

For complete summaries and access to the talks, go to **www.nyas.org/compbio**

May 11, 2005

Sequence Signals for RNA Processing

- Searching for the right words: Computational approaches to understanding RNA stability regulation, *Barrett Foat, Columbia University*
- Genetic determinism and the central dogma: The path to an RNA splicing code, *Christopher Burge, MIT*

March 9, 2005

Seeing Biochemistry: Computational methods provide microscopic visions into biological systems

- The twists and turns of lipid bilayers, *Richard Pastor, US Food and Drug Administration*
- Lipid-Protein interactions viewed through an MD lens, *Alan Grossfield, IBM T.J. Watson Research Center*
- Antifreeze proteins and the angular structure of water, *Kim Sharp, University of Pennsylvania*

February 17, 2005

Base Camp: Developing strategies for predicting microRNA targets

- Worlds apart: microRNAs and their targets in higher plants and animals, *David Bartel, MIT*
- Looking for teamwork: predicting targets for microRNA cooperativity, *Nikolaus Rajewsky, NYU*
- Learning from validated microRNA/messenger RNA targets, *Frank Slack, Yale University*

January 12, 2005

Reaching for Biology's Holy Grail: Novel methods for understanding protein sequence alignment

- Identifying co-conserved patterns in multiple-sequence alignments, *Andrew Neuwald, Cold Spring Harbor Laboratory*
- Upgrading the toolbox for protein structure prediction, *Roland L. Dunbrack, Fox Chase Cancer Center*