Biography of Phil Green

Scientific research is becoming increasingly interdisciplinary, and, consequently, areas previously thought to have little overlap are being combined to produce revolutionary advances. Two such areas, biology and math, form the basis for the unusual career of computational biologist Phil Green. A professor in the genome sciences, bioengineering, and computer sciences departments of the University of Washington (Seattle), Green has developed novel software packages for systematically analyzing complex genomes. He and his colleagues also have used computational methods to investigate a number of aspects of genome sequences, trimming down previous estimates of the number of human genes and examining the basis of molecular evolution.

Green's work has earned him numerous awards and grants, including an appointment as a Howard Hughes Medical Institute Investigator, election to the National Academy of Sciences in 2001, and a Gairdner International Award in 2002. In his Inaugural Article (1), published in this issue of PNAS, Green and his graduate student, Dick Hwang, describe a mathematical model they developed to investigate "context effects": the influence of flanking nucleotides on mutation probabilities. By applying this model, the researchers add insight to various phenomena in the molecular evolution of mammals, including the mutation rates in CpG hotspots.

Mathematical Aesthetics

Green's interest in mathematics got off to an early start when, at around age 8, he ran across a book containing a collection of number theory problems in the public library. "I just felt an immediate affinity to mathematics, and early on identified that as my main interest," he said, "but it seemed like a game." Green frequently turned to math games as a diversion from his other schoolwork and activities. "A few years later, I heard that you could actually do this for a living and got incredibly excited about it," he added.

Green's introduction to genetics came via his next door neighbor, Robert Elston, in his hometown of Chapel Hill, NC. When Green was in high school, Elston, then a statistical geneticist at the University of North Carolina (UNC) in Chapel Hill, offered him a summer job in the university's biostatistics department writing computer programs to do genetic analyses. "Robert had gone into

Phil Green

a career that involved applying mathematical ideas to genetics," said Green. "Maybe he was trying to nudge me in that direction. It didn't work immediately, since I was intent on becoming a mathematician, but working for him I got my first exposure to both computer programming and genetics."

In 1968 Green entered Harvard College (Cambridge, MA), where he majored in pure math and focused on problems in areas such as number theory and abstract algebra. After graduating in 1972, he entered the Ph.D. program in mathematics at the University of California at Berkeley, where he worked on the representation theory of operator algebras and locally compact groups under the mentorship of Marc Rieffel. This field, which combines ideas from abstract algebra and functional analysis, captivated Green both for its theoretical beauty and for its relationship to problems in physics. "It was appealing to do work that not only had aesthetic value but also a connection to the real world," he said. His thesis work resulted in two published papers (2, 3).

Green's first job after completion of his degree in 1976 was an assistant professorship in the mathematics department at Columbia University (New York). He continued his research on operator algebras, but he began to feel uncertain that this was the right career for him. "What made me increasingly dissatisfied with working in math was the sense that it was mainly aesthetically driven, at least the areas that I was working on," said Green. "The connection to the real world wasn't as strong as I had hoped."

Unusual Applicant

A turning point came for Green after he returned to Columbia after a yearlong visit to the Institute for Advanced Study in Princeton. One of his Columbia colleagues, Avner Ash, was struggling with number theory calculations too complex to solve by hand. Green wrote a program to perform the calculations on a computer. This both sped Ash's research (4) and revived Green's interest—from his time at the biostatistics department at UNC—in programming as a scientific tool. Importantly, it also gave him some ideas for guiding his career in a new direction. At about the same time, Green picked up a copy of James Watson's book *Molecular Biology of the Gene* (5). "Like many other readers, I found Watson's view of biology to be incredibly inspiring," said Green. "In some ways it was almost mathematical in its elegance. It got me thinking about genetics again."

Green decided to try to combine computer programming with genetics again, much as Elston had encouraged him not to worry about the difficulty of switching fields. "I knew that I was unlikely to be able to get somebody who didn't know me to take a chance on giving me a job while I made this career change," he recalled. To solve his dilemma, Green contacted Jim Grizzle, chair of UNC's biostatistics department, where Green had worked as a teenager. Grizzle arranged for a research associate position in the biostatistics department, and, in 1980, Green moved back to Chapel Hill to join a research team led by Kadambari Namboodiri that was analyzing blood lipid levels in families to investigate the genetic basis of heart disease. Green and his collaborators published multiple papers on this subject over the next 2 years (e.g., ref. 6).

The project made Green interested in learning more about molecular biology techniques. In the summer of 1983, he attended an intensive course in cellular and molecular biology at the Marine Biology Laboratory in Woods Hole, MA. Unlike Green, most of the classmates were graduate students in the biological sciences looking to extend their technical knowledge. Green was marked by his inexperience and atypical mathematics background. "I was later told by the course leader, Joel Rosenbaum, that they really felt they were taking a chance accepting me because I was a

This is a Biography of a recently elected member of the National Academy of Sciences to accompany the member's Inaugural Article on page 13994.

pretty unusual applicant," he said. Over the next 7 weeks, Green and his fellow students attended lectures each morning and then worked in the laboratory until past midnight performing experiments that used such techniques as Southern blots, cloning, and sequencing by nonautomated methods. "That was sort of a baptism by fire, I guess, in terms of learning experimental molecular biology," said Green.

Taking his new skills back to UNC, Green transferred to the department of pathology on a postdoctoral research grant arranged by John Graham, a genetic hematologist with whom Green had collaborated. Rather than more data analysis, this position involved mostly laboratory work. Under the guidance of junior faculty member Dana Fowlkes, Green attempted to dissect the human fibrinogen promoter by inserting gene constructs into adenovirus. The work was slow and tedious and made little contribution to science, Green recalled. However, "it was crucial to me in terms of getting experience working in the lab and understanding the many practical issues involved with real-world data," said Green. "I was getting a feel for how you design experiments. I was learning to think more like a biologist and less like a mathematician."

When his postdoctoral work concluded in 1986, Green searched for a job in which he could combine his mathematics background with his new biological knowledge. He seemed to find a perfect fit in the company Collaborative Research (CRI), based in Waltham, MA. At the instigation of scientific adviser David Botstein, CRI had undertaken a project in the early 1980s to make a genetic linkage map of the human genome based on restriction fragment length polymorphisms (RFLPs), DNA markers that could then be used to track the inheritance of genetic diseases in families and thereby localize the disease genes to specific chromosomal locations. Green joined a CRI research team led by Helen Donis-Keller, and he simultaneously began a collaboration with Eric Lander at the Whitehead Institute (Cambridge, MA) to develop the mathematical methods and computer software necessary to construct the maps (7). Whereas then-current methods allowed researchers to map three or four of these markers at a time, CRI had data on hundreds of markers, "dozens on every chromosome," said Green. He and his colleagues began using their new analysis methods to construct maps for each chromosome. In 1987, they published a paper in *Cell* detailing the first genetic linkage map of the entire human genome (8).

Back to Academics

Green found this work rewarding but in some ways not as stimulating as his former academic jobs. "One thing I definitely missed [in the private sector] was that you don't interact with as broad a variety of people with interests in different areas as you do in universities," said Green. "You don't have the option of walking across campus and talking to an algorithm specialist in the computer science department or to a biologist who is an expert in some area that you're not familiar with." Consequently, in 1989, Green left CRI for an assistant professorship in the genetics department at Washington University in St. Louis, chaired by Dan Hartl. Washington University's department of genetics was then emerging as a major center of genomics research. Maynard Olson's laboratory was developing new technologies for physical mapping, David Schlessinger was using these to make clone maps of human chromosomes, and Bob Waterston was planning with John Sulston to begin sequencing the *Caenorhabditis elegans* genome. Green began collaborations with all three groups to develop computational methods for data analysis (9, 10), but he was particularly interested in the *C. elegans* sequencing, and he wasted no time in contacting Waterston to collaborate. He was soon involved in doing mathematical calculations to estimate how much sequence data would be needed in order to have a complete genomic map.

"I was learning to think more like a biologist and less like a mathematician."

"Once the data started coming in, there was just a huge amount of interesting biology to find out from analyzing the sequence," Green said. "The mapping projects were interesting from the standpoint of computational challenges but really didn't involve any biology." Spurred by comparative genomics research taking place concurrently at other laboratories, he and his colleagues began taking sequences they had isolated from *C. elegans* and searching in the GenBank database for homologous genes conserved in widely divergent organisms, such as *Saccharomyces cerevisiae* and humans. Green's team found that nearly all the conserved genes turned out to correspond to previously studied gene families. "The conclusion seemed to be that the phylogenetically broadly distributed gene families had mostly all been discovered," he said. "It was kind of surprising, but it was also in a way reassuring that the protein universe was finite and that there didn't seem to be an infinite number of important genes to find. But, conversely, it implied that each genome seemed to have a large number of genes unique to its phylum." He and his colleagues published their findings in *Science* in 1993 (11).

In 1992, immunologist and technology developer Leroy Hood moved from the California Institute of Technology (Pasadena) to the University of Washington in Seattle to open a department of molecular biotechnology. The idea behind the department was to allow people from diverse backgrounds, including engineering, wet lab biology, protein chemistry, and computer science, to intermingle and produce revolutionary advances. "The hope was that interaction among all these people was really going to push biology to the next level," said Green. Cautious about the untried enterprise, he watched as Hood's department attracted leading researchers to its faculty, including Olson, human polymorphism expert Debbie Nickerson, protein chemists John Yates and Ruedi Aebersold, and cytogeneticist Barbara Trask. When the department reached a critical mass, Green decided to join it in 1994.

When Green arrived at University of Washington, he brought with him prototypes of two computer programs, PHRED (Phil's revised editor) and PHRAP (Phil's revised assembly program), which he and his colleagues had developed in conjunction with the *C. elegans* sequencing project at Washington University. PHRED calls bases from automated sequencer data and assigns an error probability to each base call. PHRAP assembles sequencing reads to reconstruct the underlying sequence of a genomic segment. The programs had been designed with the idea of facilitating high-throughput sequencing. "It was clear that genome sequences were really going to transform biology, so I was thinking a lot about how to accelerate the process on the data analysis side and make it as automated as possible," Green said. He and his Seattle colleagues set out to improve these two programs and to develop a third program, CONSED (consensus editor), for editing sequence data (12–14), and they began to distribute all three programs for free to academic institutions. Since they gave out the first copies in the mid-1990s, the programs "have really taken on a life of their own," Green said, in part because

sequencing has become more ubiquitous than he ever expected.

Green continues to use these programs, as well as others developed within his laboratory, to address a variety of biological questions by means of sequence analysis. A primary goal is to help develop a complete "parts list" of the proteins encoded within the genome. To that end, in 2000, he and his colleagues used ESTs to estimate how many genes exist in the human genome. Their results were surprising: although previous estimates were ≈60,000–100,000, Green's group came up with an estimate of only ≈35,000. Completion of the human genome sequence confirmed that the gene number is low. "It's still just an estimate, because no one has a complete list of the genes at this point and likely won't for several more years, but it's pretty clear that the number is probably even less than we had estimated, perhaps 30,000 genes," he said. The team published their findings in *Nature Genetics* (15).

Bringing Order

More recently, Green and his colleagues have turned to the problem of molecular evolution, particularly understanding the mutational process behind sequence changes and analyzing these changes for evidence of patterns. "I divide our interests in genome sequence interpretation into what I call signal and noise, where noise refers to the mutational process and signal is the functionally important biological information that's encoded in the genome," he explained. "Most biologists are interested in the signal. But one of our reasons in being interested in the noise is it can help you pick out the signal."

In Green's Inaugural Article (1), he and Hwang analyzed the "noise" of context effects: the influence of flanking nucleotides on the probability of a mutation happening at a particular nucleotide site. The two investigators developed a rigorous mathematical model for the context effects and applied it to characterize trends in mammalian evolution. One of their most striking findings was that mutations of the dinucleotide CpG, which are thought to arise from deamination of methylated cytosine, accumulate in a relatively clock-like fashion, in contrast to other types of mutation. Green hopes that this finding, and others not possible without such a comprehensive mathematical model, will help researchers understand the time scale of molecular mutation rates.

In the future, Green plans to stay focused on completing the genetic "parts list" and putting those findings in the context of a "wiring diagram," an explanation of how various genetic parts interact. He reconciles his dual interest in mathematics and biology as a search for order: "I actually think of mathematics at some very basic level as not all that different from biology," said Green. "In both cases, you're trying to understand some highly complex entity: an abstractly defined construct in the case of mathematics, a living organism in the case of biology. If you find some new principle involving that entity that relates several apparently unrelated observations, that's incredibly exciting—you see an order where there hadn't been any before."

Christen Brownlee, *Science Writer*

1. Hwang, D. G. & Green, P. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 13994–14001.
2. Green, P. (1977) *Pac. J. Math.* **72,** 71–97.
3. Green, P. (1978) *Acta Math.* **140,** 191–250.
4. Ash, A., Grayson, D. & Green, P. (1984) *J. Number Theory* **19,** 412–436.
5. Watson, J. D. (1976) *Molecular Biology of the Gene* (Benjamin, New York), 3rd ed.
6. Green, P., Owen, A. R. G., Namboodiri, K., Hewitt, D., Williams, L. R. & Elston, R. C. (1984) *Genet. Epidemiol.* **1,** 123–141.
7. Lander, E. & Green, P. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2363–2367.
8. Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T., Bowden, D., Smith, D., Lander, E., *et al*. (1987) *Cell* **51,** 319–337.
9. Green, E. D. & Green, P. (1991) *PCR Methods Appl.* **1,** 77–90.
10. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., *et al*. *Nature* (1992) **356,** 37–41.
11. Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. & Claverie, J.-M. (1993) *Science* **259,** 1711–1716.
12. Ewing, B., Hillier, L., Wendl, M. & Green, P. (1998) *Genome Res.* **8,** 175–185.
13. Ewing, B. & Green, P. (1998) *Genome Res.* **8,** 186–194.
14. Gordon, D., Abajian, C. & Green, P. (1998) *Genome Res.* **8,** 195–202.
15. Ewing, B. & Green, P. (2000) *Nat. Genet.* **25,** 232–234.