

FEBRUARY 12, 2001

Comparing Genomes Will Illuminate Gene Function

The surprising similarities and startling differences between the genomes of humans, fruitflies, worms, plants and yeast constitute a scientific treasure trove that may lead to unprecedented insight into how genes function, according to Gerald M. Rubin of the Howard Hughes Medical Institute (HHMI).

In a *News and Views* article published in the February 15, 2001, issue of the journal *Nature*, Rubin, HHMI's vice president for biomedical research and leader of the Berkeley *Drosophila* Genome Project, outlined some of the benefits that might be obtained by comparing the genomes of vertebrates, invertebrates and plants. The article is part of a collection of articles published by *Nature* that discusses the implications of sequencing the human genome.

"Over most of my career, people could plan their experiments over a weekend, spend six months doing them, and then interpret the results over a weekend. Now, people can do an experiment over a weekend and spend six months thinking about what the results mean."

— Gerald M. Rubin

The obvious starting point for comparing the genomes of different species, writes Rubin, is with the total number of genes. "Here is a real surprise: the human genome probably contains between 25,000 and 40,000 genes, only about twice the number needed to make a fruitfly, worm or plant," Rubin wrote. Some increased complexity in vertebrates is achieved by more frequent use of "alternative splicing," where protein-coding segments of genes are spliced in such a way that a single gene may yield a number of messenger RNA molecules, and thereby produce a greater variety of proteins. However, this does not appear sufficient to account for the large apparent differences in complexity between species, Rubin said.

In an interview discussing the *Nature* article, Rubin said, "We first noted this lack of correlation between gene number and complexity when we found that a fruitfly has only about twice as many genes as yeast. Yet, a fruitfly is a complicated animal, with complex behaviors and circadian rhythms. It can fly around and not crash into walls and seems much more than twice as complex as a unicellular yeast.

"A good analogy is that these genes are like a set of LEGOs, the children's blocks. With the same simple set of units, you could build a complete scale model of the Vatican or you could build a log cabin," said Rubin.

"For example, a human brain is much more complex than a fruitfly brain. But if you compare the individual nerve cell in the fruitfly to a human nerve cell, they're really not that different. The human brain is much more capable because it has orders of magnitude more cells, interconnected in very complex ways," said Rubin.

Writing in *Nature*, Rubin also noted that 90 percent of structural units, or domains, that can be identified in human proteins, are also found in fruitfly and worm proteins. Despite this fact, more than a third of the proteins in yeast, fruitflies, worms and humans show no strong similarity across species. These proteins might have similar functions but different structure; or they might show species-specific functions, Rubin wrote. Or, they might have an unknown evolutionary mechanism for maintaining themselves that is independent of their precise sequence.

"We have found lots of surprises in biology over the years," he said. "And there are likely many new ones still lurking."

Rubin maintains that one of the key approaches to understanding the function of individual DNA segments will be to compare genes of closely related species. "The concept is simple: segments that have a function are more likely to retain their sequence during evolution than non-functional segments," he wrote. "So DNA segments that are conserved between species are likely to have important functions."

"This is sort of nature's way of doing the same kinds of experiments that scientists do in the laboratory," he said in an interview. "Just as researchers alter genes and explore the changes in function, so evolution has done the same thing, and all we have to do is analyze the effects."

But understanding how gene expression is regulated is much more difficult than analyzing gene structure, wrote Rubin. Even though regulatory regions of genes have been identified, he wrote, "the proteins that control gene expression by recognizing regulatory regions often detect sequence features that elude the best computer algorithms, and may use information from contacts with other proteins that is difficult to model.

"Proteins are simply cleverer than computers," he wrote. However, he expressed confidence that such problems will be solved.

"As methods for comparing sequences continue to improve," he wrote, "we can expect to learn more about elusive features of the genome, such as genes encoding RNAs that do not encode proteins, start points of DNA replication, and genetic elements that control chromosome structure."

Rubin commented that "people are beginning to find that when they compare closely related species, they find pieces of DNA that are conserved in evolution and therefore they assume have some function. Soon we'll have a list of these conserved pieces, and a long list of functions we know exist; and we'll begin figuring out how these two lists are correlated."

Rubin said that progress will accelerate as researchers learn to cope with the vast quantities of data being generated by genomic research. "Over most of my career, people could plan their experiments over a weekend, spend six months doing them, and then interpret the results over a weekend. Now, people can do an experiment over a weekend and spend six months thinking about what the results mean," he said. "While we are reaching a barrier in terms of just using the unaided human brain to interpret the data, I would rather call it an opportunity for people to develop new approaches," he said.