

JANUARY 28, 2005

How Many Comparative Genomes Are Enough?

As the human genome sequence neared completion several years ago, geneticists eagerly began discussing which other organisms to sequence—partly to see which DNA regions are similar across species and therefore likely to serve critical functions. But these discussions raised an important, and potentially expensive, question: How many species need to be sequenced to know whether evolution has conserved a given stretch of DNA?

In an article published in the January 2005 issue of *PLoS Biology*, Sean R. Eddy, a Howard Hughes Medical Institute investigator at Washington University School of Medicine in St. Louis, describes a mathematical model that offers detailed answers to this question. “We shouldn't make these decisions based on seat-of-the-pants intuitions,” Eddy said. “It's important to lay out the case that these genomes really do have tremendous value for analyzing the human genome sequence.”

According to Eddy's model, critical tradeoffs are associated with deciding which species to sequence. More species need to be compared to tell if just one or a few DNA bases are conserved, compared to what is needed to identify longer stretches of conserved DNA. Also, the more closely related a group of organisms is in evolutionary terms, the more comparisons need to be made to tell if a given DNA region is conserved across species.

"We shouldn't make these decisions based on seat-of-the-pants intuitions. It's important to lay out the case that these genomes really do have tremendous value for analyzing the human genome sequence."

- Sean R. Eddy

Previous analyses had indicated that 10 to 20 well-chosen mammalian genomes would be enough to determine whether any given nucleotide is conserved with an error rate of less than 1 in 100. But this estimate does not

take into account that nucleotides need not remain exactly the same over time to be conserved between species, Eddy points out. Furthermore, deciding that a single base is conserved is not necessarily the most appropriate goal. According to Eddy, it would be more useful to figure out if longer stretches of DNA are conserved, such as sections of genes or binding sites in DNA for proteins that control gene expression.

Eddy constructed a mathematical model that takes into account the length of a conserved DNA region, the number of different species sequenced, and their evolutionary distance. His model assumes that conserved nucleotides do not necessarily stay the same but evolve at a slower rate than nucleotides that are not conserved.

The model finds, in accordance with previous results, that detecting invariant single nucleotides would require comparing about 17 genomes separated by the average evolutionary distance between humans and mice. But when conserved nucleotides are allowed to change in a more realistic way, 25 genomes are needed instead. To reduce the error rate from 1 in 100 to 1 in 10,000, about 120 such genomes should be compared.

However, far fewer genomes are needed to detect conserved features larger than a single nucleotide. For parts of genes about 50 nucleotides long, only a single comparison is needed, and even for gene segments eight or so nucleotides long (such as binding sites for transcription factors), 3 to 15 genomes are needed.

The model holds up well when applied to simulations of realistic nucleotide evolution patterns. It also accurately predicts results derived from existing genome comparisons.

“No one had actually written out the case for why we are proposing to sequence the koala and the bat and the platypus,” Eddy said. “This is one way of showing that, yes, you need a fair amount of statistical information from these comparative genomes.”

Eddy “did a nice job of making the intuitive rigorous,” said Philip Green, an HHMI investigator at the University of Washington in Seattle. His work “will be useful in guiding people's thinking about how to use comparative data and how much data you need.”

Eddy took the unusual step of publishing a conflict-of-interest statement in his article in *PLoS Biology* that notes that he is associated with a genome sequencing center. “I was very conscious of conflict of interest considerations,” he said. “Here I am sitting at a genome sequencing center saying that we need 100 genomes, not one. . . . But it's important for programmatic decisions that we try to do more modeling and pilot studies so that we can justify these millions of dollars spent on sequencing.”