

AUGUST 03, 2004

## Using Statistics to Decipher Secrets of Natural Mutation

A new mathematical approach for analyzing the complex, subtle patterns of natural mutation in DNA is likely to help biologists understand how mutation contributes to evolutionary change in mammals.

The researchers, Howard Hughes Medical Institute investigator Philip Green and his student Dick Hwang, published a report describing the first applications of their new analytical approach in the August 3, 2004, online early edition of the *Proceedings of the National Academy of Sciences*. Both Hwang and Green are at the University of Washington in Seattle.

---

"Understanding naturally occurring mutations has been of great interest because mutations are major drivers of evolution."

— Philip Green

---

"Understanding naturally occurring mutations has been of great interest because mutations are major drivers of evolution," said Green. "However, it's surprising how little is still known about their causes."

Previous studies have revealed a number of biases in the rates of different types of mutational change. These arise in part from the innate biochemical characteristics of the four DNA nucleotide bases - adenine, guanine, cytosine and thymine - that affect their vulnerability to modification and the accuracy with which they are replicated when cells divide. Particular nucleotide sequences, for example, cytosine-guanine (CpG) dinucleotides, form "hotspots" - regions that are particularly vulnerable to alterations that convert one nucleotide to another, causing mutations.

To understand these biases, Hwang and Green sought to develop a flexible approach to analyze the process of "neutral DNA evolution" in regions of the genome thought to lack genes and other functionally important sequences. "If you want to get an unvarnished picture of the mutation process itself, uncorrupted by natural selection, you want to look at neutrally evolving DNA," said Green. "Mutations in DNA that is not functional should better represent the complete spectrum of naturally occurring mutations. Mutations are of course also occurring in the genes and those are of interest because

they can create new phenotypes and cause variation among traits. Some of those mutations are advantageous and consequently quickly spread through the species, while others are deleterious and are weeded out. So genes and other features don't evolve at neutral rates.”

“Apart from their intrinsic interest, we think understanding the underlying mutation patterns better will also help us in finding the functionally important features in the genome. Basically, it's a signal-to-noise issue, where the naturally occurring mutations are the 'noise' and the functional parts of the genome are the 'signal.' The better we understand the noise, the better job we can do of understanding the signals.”

To begin to understand the patterns of DNA changes that result from neutral mutation, Hwang and Green developed a new version of a powerful statistical technique that they call “Bayesian Markov chain Monte Carlo sequence analysis.” Basically, the technique enables them to feed in sequence information from genomes of different organisms and discern patterns that can distinguish models of mutational mechanisms.

According to Green, the statistical approach offers an effective way to analyze models that are very difficult or impossible to solve analytically. “Until recently,” he said, “the state of the art in the molecular evolution field was to use models that people knew were gross over-simplifications, but had the merit that you could solve them analytically. Without doing too much computation, you could make estimates of mutation rates of various sorts. However, the cost of that simplified approach was a model that is unrealistic.”

In particular, he said, the standard model treated all positions in the sequence as evolving independently of each other, rather than taking into account context effects, in which the identity of neighboring nucleotides influences the nature and rate of mutations.

“While a few other investigators have been working on how to take into account context effects, I think we are doing it in a more rigorous, more complete way,” said Green. Without such a rigorous approach, he said, models of evolution could give erroneous results regarding the effects of mutation.

“I think the more realistic you can make the model, the less likely you are to be led astray by drawing conclusions that really had more to do with the deficiencies of your model than with the underlying reality,” said Green.

Hwang and Green tested their analytical approach by using it to compare the sequences of corresponding genome segments from 19 mammalian species, including human, horse, lemur, rat, rabbit, hedgehog and armadillo. Such comparisons among species across the mammalian evolutionary tree can provide insight into how mutational patterns have changed over evolutionary time.

They focused their analysis on a 1.7 million base-pair DNA segment known as the “greater cystic fibrosis transmembrane conductance regulator region,” which was sequenced in the 19 mammals by Eric Green and his colleagues at the National Human Genome Research Institute. To concentrate on the neutrally evolving DNA, Hwang and Phil Green excluded the genes from those segments and compared what was left.

According to Green, the comparison of context-dependent mutation in the segments across the species revealed that the CpG mutations, unlike other mutation types, accumulated in a regular clock-like fashion. The analysis also distinguished other sources of naturally occurring mutations and their variation due to biological and biochemical influences, and appears to offer some insight into factors such as generation time and population size that have varied in mammalian evolution.

Green concluded that by contributing to a better understanding of naturally occurring mutations, the technique would help in understanding both how genetic disease arises and how evolution has occurred.

A next step, he said, will be to extend the analysis to sites on the genome that are not evolving neutrally. This should help identify genomic regions that were not previously recognized to be of functional importance, said Green. Also, he said, such analyses could offer considerable insight into how patterns of natural selection have varied across different species in the course of evolution.

“A more complex model of the neutral process should start to pay for itself in exploring these phenomena, because you're frequently looking for relatively subtle effects,” said Green.