Taking the gamble out of DNA sequencing

February 24 2013

Two USC scientists have developed an algorithm that could help make DNA sequencing affordable enough for clinics – and could be useful to researchers of all stripes.

Andrew Smith, a computational biologist at the USC Dornsife College of Letters, Arts and Sciences, developed the algorithm along with USC graduate student Timothy Daley to help predict the value of sequencing more DNA, to be published in Nature Methods on February 24.

Extracting information from the DNA means deciding how much to sequence: sequencing too little and you may not get the answers you are looking for, but sequence too much and you will waste both time and money. That expensive gamble is a big part of what keeps DNA sequencing out of the hands of clinicians. But not for long, according to Smith.

"It seems likely that some clinical applications of DNA sequencing will become routine in the next five to 10 years," Smith said. "For example, diagnostic sequencing to understand the properties of a tumor will be much more effective if the right mathematical methods are in place."

The beauty of Smith and Daley's algorithm, which predicts the size and composition of an unseen population based on a small sample, lies in its broad applicability.

"This is one of those great instances where a specific challenge in our research led us to uncover a powerful algorithm that has surprisingly

broad applications," Smith said.

Think of it: how often do scientists need to predict what they haven't seen based on what they have? Public health officials could use the algorithm to estimate the population of HIV positive individuals; astronomers could use it to determine how many exoplanets exist in our galaxy based on the ones they have already discovered; and biologists could use it to estimate the diversity of antibodies in an individual.

The mathematical underpinnings of the algorithm rely on a model of sampling from ecology known as capture-recapture. In this model, individuals are captured and tagged so that a recapture of the same individual will be known – and the number of times each individual was captured can be used to make inferences about the population as a whole.

In this way scientists can estimate, for example, the number of gorillas remaining in the wild. In DNA sequencing, the individuals are the various different genomic molecules in a sample. However, the mathematical models used for counting gorillas don't work on the scale of DNA sequencing.

"The basic model has been known for decades, but the way it has been used makes it highly unstable in most applications. We took a different approach that depends on lots of computing power and seems to work best in large-scale applications like modern DNA sequencing," Daley said.

Scientists faced a similar problem in the early days of the human genome sequencing project. A mathematical solution was provided by Michael Waterman of USC, in 1988, which found widespread use. Recent advances in sequencing technology, however, require thinking differently about the mathematical properties of DNA sequencing data.

"Huge data sets required a novel approach. I'm very please it was developed here at USC," said Waterman.

Provided by University of Southern California