Medical Xpress

Finding deep roots, new genome software infers ancestry with high accuracy

March 19 2008

Some people may know where their ancestors lived 10 or 20 generations ago, but the rest of us can learn our distant biological heritage only from our DNA. New genomics analysis software developed by computer scientists at Stanford appears far more adept than prior methods at unraveling the ancestry of individuals. A paper describing the HAPAA system, which takes its name from "hapa," the Hawaiian word for someone of mixed ancestry, appears online today and in the April printed issue of the journal *Genome Research*.

Going back 20 generations the software can identify what continent or broad global region an individual's ancestors were from. But going back about 10 generations the software can be much more precise, making distinctions as fine-grained as the traditional gene pools of nearby population groups—hypothetically differentiating Greek from Italian, or Russian from German.

Specifically what the software does is compare an individual to all those in the International HapMap database to see what distinct spans of genetic snippets, called haploblocks, they share in common.

"With very high accuracy, even for 20 generations, we can trace the populations of those individuals who are indeed represented in your genome," says Stanford computer science Assistant Professor Serafim Batzoglou, who led a team of graduate students to create HAPAA. They include co-lead authors Andreas Sundquist and Eugene Fratkin, as well as Chuong B. Do.

Batzoglou points out that because the HapMap database, a genetic record of 270 individuals of Western European, West African and East Asian ancestry, is very small, HAPAA now can only generate an ethnic profile in terms of these populations.

Fratkin himself was able to verify that he is of European ancestry, but not that he is 1/64th Polish. But more genomics data will become available, the researchers said, which will further expand the software's ability to help people discern their roots.

Low error, high precision

In the Genome Research paper the researchers tested the system's accuracy using real individuals in the database and by synthesizing virtual people, essentially simulating mating for 20 generations among individuals in the database.

The team also compared HAPAA to the current state-of-the-art system known as SABER. Using the standard statistical measure of "mean-square" error, Batzoglou and his students found that HAPAA's error rates were between a half and a third as big as SABER's. The difference widened as the generations probed went further back—meaning that HAPAA's error rate remains consistently low, even back 15 or 20 generations.

An important advance that improves HAPAA's accuracy is its more accurate modeling of individual variation. The Stanford computer scientists created an algorithm efficient enough to compare the genetic information of the test individual to that of every individual in the database. Other systems, including SABER, rely on comparisons to a composite that represents an averaging of the data from many individuals. That methodology is easier to program and run on a computer, but the problem with averaging is that a lot of information is lost.

Consider using comparison as the way to characterize a soccer player. One could look at her total goals scored and compare that figure to historical league average. Such a comparison would reveal whether she was generally a high scorer, but couldn't lend any insight as to whether her scoring patterns (e.g., game winners, late-game goals, penalty kicks) were more like those of Mia Hamm or Birgit Prinz.

For now the HAPAA software provides proof of this concept but limited utility given the small size of the HapMap database. In the future the software will benefit not only from having more individuals available for comparison, Batzoglou said, but also more detailed data about each individual. Today's genome samples track about 500,000 markers, or common genetic differences, but there are about 10 million candidates. Most individuals have about 3 million such specific differences. As genomics technology improves, he says, so will HAPAA's ability to infer ancestry from the data.

The research was supported by a grant from the National Institutes of Health and a Stanford graduate fellowship provided by the German software company SAP AG.

Source: by David Orenstein, Stanford University

Citation: Finding deep roots, new genome software infers ancestry with high accuracy (2008, March 19) retrieved 20 April 2024 from https://medicalxpress.com/news/2008-03-deep-roots-genome-software-infers.html