PHYS ORG

Searching genomic data faster with new algorithm

July 10 2012, by Larry Hardesty

In 2001, the Human Genome Project and Celera Genomics announced that after 10 years of work at a cost of some $400 million, they had completed a draft sequence of the human genome. Today, sequencing a human genome is something that a single researcher can do in a couple of weeks for less than $10,000.

Since 2002, the rate at which genomes can be sequenced has been doubling every four months or so, whereas computing power doubles only every 18 months. Without the advent of new analytic tools, biologists' ability to generate genomic data will soon outstrip their ability to do anything useful with it.

In the latest issue of Nature Biotechnology, MIT and Harvard University researchers describe a new algorithm that drastically reduces the time it takes to find a particular gene sequence in a database of genomes. Moreover, the more genomes it's searching, the greater the speedup it affords, so its advantages will only compound as more data is generated.

In some sense, this is a data-compression algorithm — like the one that allows computer users to compress data files into smaller zip files. "You have all this data, and clearly, if you want to store it, what people would naturally do is compress it," says Bonnie Berger, a professor of applied math and computer science at MIT and senior author on the paper. "The problem is that eventually you have to look at it, so you have to decompress it to look at it. But our insight is that if you compress the data in the right way, then you can do your analysis directly on the compressed data. And that increases the speed while maintaining the accuracy of the analyses."

Exploiting redundancy

The researchers' compression scheme exploits the fact that evolution is stingy with good designs. There's a great deal of overlap in the genomes of closely related species, and some overlap even in the genomes of distantly related species: That's why experiments performed on yeast cells can tell us something about human drug reactions.

Berger; her former grad student Michael Baym PhD '09, who's now a visiting scholar in the MIT math department and a postdoc in systems biology at Harvard Medical School; and her current grad student Po-Ru Loh developed a way to mathematically represent the genomes of different species — or of different individuals within a species — such that the overlapping data is stored only once. A search of multiple genomes can thus concentrate on their differences, saving time.

"If I want to run a computation on my genome, it takes a certain amount of time," Baym explains. "If I then want to run the same computation on your genome, the fact that we're so similar means that I've already done most of the work."

In experiments on a database of 36 yeast genomes, the researchers compared their algorithm to one called BLAST, for Basic Local Alignment Search Tool, one of the most commonly used genomic-search algorithms in biology. In a search for a particular genetic sequence in only 10 of the yeast genomes, the new algorithm was twice as fast as BLAST; but in a search of all 36 genomes, it was four times as fast. That discrepancy will only increase as genomic databases grow larger, Berger explains.

Matchmaking

The new algorithm would be useful in any application where the central question is, as Baym puts it: "I have a sequence; what is it similar to?" Identifying microbes is one example. The new algorithm could help clinicians determine causes of infections, or it could help biologists characterize "microbiomes," collections of microbes found in animal tissue or particular microenvironments; variations in the human microbiome have been implicated in a range of medical conditions. It could be used to characterize the microbes in particularly fertile or infertile soil, and it could even be used in forensics, to determine the geographical origins of physical evidence by its microbial signatures.

Berger's group is currently working to extend the technique to information on proteins and RNA sequences, where it could pay even bigger dividends. Now that the human genome has been mapped, the major questions in biology are what genes are active when, and how the proteins they code for interact. Searches of large databases of biological information are crucial to answering both questions.

Provided by Massachusetts Institute of Technology

Citation: Searching genomic data faster with new algorithm (2012, July 10) retrieved 25 April 2024 from https://phys.org/news/2012-07-genomic-faster-algorithm.html