Program turns up 150 missed genes

May 11 2005

A computer scientist at Washington University in St. Louis has applied software that he has developed to the genome of a worm and has found 150 genes that were missed by previous genome analysis methods. Moreover, using the software, he and his colleagues have developed predictions for the existence of a whopping 1, 119 more genes.

Image: This is C. elegans. Its genome was thought to have been completed until a WUSTL computer scientist applied a computer software program he developed which found scores more genes and predicted the existence of over a thousand additional genes.

Michael Brent, Ph.D., Washington University professor of computer science and engineering, used his unique software, TWINSCAN, on the genome of Caenorhabditis elegans (C. elegans). The genome of another nematode C. briggsae, was also used to determine which parts of the sequence have changed since the nearest common ancestor of the two species. He found first of all that TWINSCAN predicted 60 percent of the genes in the C. elegans genome exactly, right own to the last amino acid.

"This (60 percent) is a new level of accuracy for a complex genome," Brent said. "It's quite a step up from what you see in the human genome, for instance, where not even a third of the genome can be predicted exactly. The 60 percent is the highest accuracy published for a multicellular organism."

C. elegans is a biological model for animal development and genetics, and is the first animal genome to be sequenced, back in 1998. Nematode researchers rely on a genome annotation database called WormBase. Along with confirmed genes, WormBase includes thousands of predicted genes without evidence from complementary DNA (cDNA) or expressed sequence tags (EST), which help locate genes. These predicted genes are derived by a combination of a program from the previous generation and some curation by human experts. Brent and his colleagues say that the accuracy of WormBase can be improved with the use of TWINSCAN predictions. And Brent predicts that the age of the human genome annotator is passing — the future belongs to computer-driven annotation.

Crossing the tipping point

"We've crossed the tipping point with gene prediction where it's becoming clear that machines can beat human annotators and analysts, on average," he said.

Because of the increasing speed of computers, the TWINSCAN analysis of C. Elegans is able to use more accurate models of intron length than previous analyses. This is important for finding exons, which house the coding machinery of proteins. While getting intron length is helpful for gene annotation, the process is 15 times slower than the typical, less accurate methods. Being able to define intron length has implications for the human genome, which is much larger than C. elegans and has an average intron length of about 4,000 base pairs, compared with an average intron length of a couple hundred base pairs in C. elegans.

Brent and colleagues from the Dana-Farber Cancer Institute and Harvard Medical School published their findings in the April, 2005 issue of Genome Research. Brent's graduate student, Chaochun Wei, is first author on the paper. The research was supported by grants from NIH, NSF, the National Cancer Institute, the National Human Genome Research Institute, and the National Institute of General Medical Sciences.

Brent has brought his bioinformatics skills to many genomes, including those of mammals, other nematode species and most recently the fungus Cryptococcus neoformans. Brent's approach to gene prediction stands traditional genome annotation on its head because it starts with a computer analysis of the genome sequence, using that as a hypothesis designing experiments to test the hypothesis. The traditional modus operandi is a data-driven approach that starts with sequencing a random sample of tens of thousands of cDNA clones. Whereas the traditional approach leads to sequencing some genes thousands of times and others not at all, Brent's approach is to sequence each predicted gene once.

"I've been building a case that we should start with predictions," he said. "Each gene sequence is more expensive, but because of the lower redundancy you end up with much better coverage of the genome for the same money."

Chess as metaphor

Brent said that some genome researchers have been reluctant to go towards an automated, hypothesis - driven approach because of a lingering sense that anything that's been looked at by a human will be more accurate than something produced by a machine.

"But look at the world of chess. Fundamentally, humans are better than machines at chess, but if you get a team of ten people with enough expertise, money, and equipment, and the willingness to work for ten years and burn a lot of computer power, they'll come up with a machine that can beat the world champion. The same principal applies to developing a machine that can reveal the mysteries of our genes. In this case, the necessary investments have been made, but since there is no sanctioned world championship, it is not yet widely known."

Source: Washington University in St. Louis

Citation: Program turns up 150 missed genes (2005, May 11) retrieved 26 April 2024 from https://phys.org/news/2005-05-genes.html