A genomic CluE for cloud computing

April 23 2009

DNA sequencing is the next frontier in biological research. As new sequencing technology becomes more efficient and affordable, it is increasingly available to small laboratories. Thus, sequencing data is being generated at a faster rate than ever before.

However, the computing capacity needed to analyze such vast amounts of data still has some catching up to do. Large networks of interconnected computers, called computer clusters, are required to analyze these data. Expensive to establish and maintain, these computer clusters are generally available only to labs that can afford them.

Enter Mihai Pop, an assistant professor in the department of computer science and in the Center for Bioinformatics and Computational Biology at the University of Maryland. He and colleague Steven Salzberg, director of the center and Horvitz Professor of computer science, recently received a grant from the National Science Foundation Cluster Exploratory Program (CluE) to fund research aimed at discovering how remote cluster computers, computer networks available over the internet, might be used to process DNA sequence data.

"There is a new initiative by NSF to figure out what you can do with cluster computers on the internet - like the ones through Amazon, Google, and IBM," Pop said. "Our NSF grant will be used to find out if remote clusters of computers are a better option for DNA sequence analysis than local clusters of computers."

Pop's goal is to develop the software required to analyze sequence data

in parallel (on many computers simultaneously). This massively parallel computing allows faster gene sequence alignment and genome assembly.

While parallel computing is already being used on locally maintained computer clusters, Pop will be working on programs that will allow researchers to perform their DNA sequence over the web by accessing remote computer clusters maintained by large companies on a pay-per-use basis. This paradigm is known as Cloud Computing.

So now, rather than buying and maintaining their own computer systems, researchers may simply be able to rent computer time at a fraction of the cost. But there are a few obstacles to overcome before Cloud Computing becomes a reality for genetic analysts.

"The first question is how to best split up the process of DNA sequence analysis to fit these computer clusters," Pop said. "The second is whether or not the benefits of cloud computing outweigh the costs of data transfer and storage."

The massive amounts of data generated by just one genome may take a significant amount of time to transfer over the internet. This, in addition to the data storage needed before analysis, might add costs that outweigh the benefits of using a remote computer cluster.

"Even if the analysis doesn't take long, the transfer may take forever and cost too much to make whole thing worthwhile," said Pop.

A Different Kind of Puzzle

DNA is made up of nucleotide bases that are abbreviated by the letters A, C, G, and T. Lined up in a double helix structure, they make up a code that is translated into the proteins that run our body processes. New technology can read this code and compare the genetic makeup of

species and organisms.

However, the sequencing process cannot handle a whole genome at once. The DNA strands have to be chopped into small pieces, sequenced, and then those sequences have to be put back together again. Putting the pieces back together is what requires so much computing power.

There are two ways to put the pieces back together. If a reference genome is available from the same species, scientists can use the reference as a guide for piecing together the new sequence. However, if a reference is unavailable, the scientist faces the more difficult task of determining all possible combinations of the loosely fitting pieces and finding the best one.

Pop likens this process to completing a jigsaw puzzle. "If you have a reference genome, it's like having the box with the picture on the front to guide your assembly," he said. "With no reference, it's like having no picture and no idea what the finished product will look like; with lots of sky and ocean pieces that fit very loosely together."

Such a process requires a lot of computing power because of the number of possibilities and level of uncertainty. Computer clusters can do all the comparisons of sequence combinations and decide on the best one. But computer power and expense of systems are a limiting factor.

Pop's team will spend the next two years determining whether it is feasible and beneficial to do this analysis through cluster computers available on the internet. He will write software programs that, if successful, will be made available for researchers to use at no cost, and his results will be made available through journal articles and conference presentations.

Teaching and mentoring of both grads and undergrads will also be a

large component of the grant, which Pop hopes will help entice talented computer science students to go into the biotechnology industry where their skills are needed.

Source: University of Maryland (news : web)