New Stanford tool enables wider analyses of genome 'deep sequencing'

May 2 2010

Life is almost unbearably complex. Humans and mice, frogs and flies toggle genes on and off in dizzying combinations and sequences during their relentless march from embryo to death. Now scientists seeking to understand the machinations of the proteins behind the genomic wizard's screen have a powerful new tool at their disposal, courtesy of researchers at the Stanford University School of Medicine.

Until now, researchers have relied on outdated methods of analysis to identify those DNA sequences involved in controlling when and how individual genes are expressed. Most often, those methods - capable of probing only specific, limited regions of the genome arising from a type of experiment called DNA microarrays - led to the exclusive scrutiny of regions called promoters nestled near the start of the gene.

In contrast, the new Stanford-developed, web-based algorithm allows scientists to plumb the unprecedented depths of the data provided by new "deep-sequencing" techniques to reveal a pantheon of control regions for nearly any gene. The effect is like expanding a researcher's field of vision from a pencil-thin beam of light trained mainly on the regions near coding sequences to a sweeping spotlight illuminating the contributions of distant genomic regions.

"It used to be that people thought only the regions near the gene were important in controlling its function - in part because they had no way of assessing the impact of regions further away," said Gill Bejerano, PhD, assistant professor of developmental biology and of computer science at

the medical school and Stanford's School of Engineering.

As a result, said Bejerano, researchers often cherry-picked nearby regions for further analysis based on their proximity or interest. "But when you're being that conservative with current sequencing capabilities, you're typically throwing away at least half of the data you so laboriously worked to obtain," he said.

Typically that data exists in the form of DNA binding sites for regulatory proteins called transcription factors that dictate the activity of genes. And, with the advent of new, deep-sequencing techniques, it's being generated at rates that are both unimaginable and unmanageable.

Bejerano is the senior author of the research, which will be published online May 2 in *Nature Biotechnology*. The researchers coined the name "GREAT" for their algorithm, an acronym for "Genomic Regions Enrichment of Annotations Tool," and the website will be available for anyone to use after May 2 at http://great.stanford.edu

There are hundreds of known transcription factors. Each controls the expression of numerous genes by binding to specific regions in the genome. This makes it difficult for scientists to know exactly how any one transcription factor is acting, particularly if it works over long stretches of DNA. Usually they'll figure out where in the DNA the protein is binding and then look for interesting genes nearby. Or, conversely, they'll find an interesting gene and look for nearby transcription-factor binding sites. But recent research has shown that sections of DNA far away can also play an important role.

It works a bit like this: Think of your kitchen. Notice all the black things. Those are your transcription-factor binding sites. But what do they do? You might figure out that sliding the lever on the toaster makes the toast pop up. And plugging it into the wall makes it get hot. But

you're likely to overlook that vitally important black breaker switch on the wall behind you, or to dismiss it as inconsequential among all the other black items in the room that don't, in fact, control the toaster. That is, unless you use this new analysis.

In contrast, users of the GREAT algorithm, developed by graduate students Cory McLean and Aaron Wenger and software engineer Dave Bristor, will simply enter a list of all the binding sites they've found throughout the genome for their transcription factor of interest. No prescreening is necessary, and the list can be hundreds or thousands of items long. Some will be biologically meaningful, and some will be experimental flukes. The software program will then provide an analysis revealing not only which genes that transcription factor is likely to moderate, both near and far, but also in which developmental or molecular pathways it is likely to function.

"The analysis gets pushed back into the hands of the person who did the experiment," said Bejerano. "Now you will start to see the kinds of results that we had expected with this much data." He and his collaborators found that test runs with well-known transcription factors verified the factors' association with the expression of particular genes, but also identified new, previously unsuspected alliances between binding sites and genes separated on the DNA by up to 1 million nucleotides.

"We've been asking the right questions, but using the wrong interpretation tools to answer them," said Bejerano. "We don't expect that this tool will help three labs. We expect that it will help 3,000 labs. GREAT can look at thousands of binding sites and tell you things that your transcription factor is doing that have never been reported before."

Provided by Stanford University Medical Center