PHYS.ORG

# Learning the language of gene expression

January 19 2007

Researchers have taken a major step towards understanding the language of gene regulation in the fruitfly Drosophila and they expect the technique to be rapidly applicable to understanding the effects of genome variation in humans.

The new research, published today in *PLoS Computational Biology*, is a major advance in using computers to detect the regions in DNA that control the activity of genes. Studies on single genes have shown that variation in gene regulation can be important in disease. The new program, called NestedMICA, allows researchers to find many regulatory regions, which will become a new focus for disease understanding.

The team, from the Wellcome Trust Sanger Institute and The University of Manchester, took slices of genome sequence from next to each Drosophila gene - where the highest concentration of regulatory signals are thought to lie - and fed them into the new computer program that looks for patterns shared between the sequences. The search process is similar to looking for words in a sentence where the vocabulary of the language is unknown

"Most words in the language of gene regulation can be spelled more than one way," explained Dr Thomas Down, first author on the report. "In English, you might see people writing either 'analyse' or 'analyze'. In genomes, such variation - or even bigger differences - seems to be normal.

"So we can't just count words, we need to recognize alternative spellings."

The team, which includes Dr Casey Bergman from Manchester's Faculty of Life Sciences, has so far found 120 'words' - distinct examples of regions that might regulate genes. About 30 of these were known from many years of studying how individual Drosophila genes are controlled, but most are novel. This is a major step towards understanding the language of gene regulation in an important model organism, and proof of principle of a new technology that will speed the study of regulatory elements in the human genome. Drosophila is a well-studied organism and shares 48% of its 14,000 genes with humans.

Research emerging in the past few months suggests that variation in the sequence of regulatory regions will affect susceptibility to many diseases. A few cases are already known - one form of thalassaemia is caused by a regulatory sequence variant - but knowledge of regulatory elements in the human genome is limited: scientists have only scratched the surface.

Systematic annotation of regulatory regions in the human genome will be very important if researchers are going to understand the effects of all sequence variation.

Dr Tim Hubbard, senior author on the report explained: "While others have tried to identify these control regions before, they have had to try to align lots of sequences. Our new method doesn't depend on alignment, an advantage because the new program is robust to rapidly evolving sequences.

"The new method also doesn't require prior knowledge from, say, looking at known examples, and can search for hundreds of different motifs at once."

As science should, the work makes predictions that the team is testing. Using a set of excellent, publicly available data on gene activity from the University of California-Berkeley and Lawrence Berkeley National Laboratory, they have predicted what some of the newly discovered sequences might mean in the language of gene regulation.

Computer analysis can accelerate the search for important regions in genomes, but the authors emphasize that computer predictions must always be examined experimentally. The findings in Drosophila by the new program have been validated by examining findings against results from experimental imaging.

The results of the research, a set of Drosophila sequence motifs, are freely available from a database at the Sanger Institute. Like many tools developed at the Sanger Institute, NestedMICA is open source software, freely available for anyone to download, run and modify.

Source: University of Manchester

Citation: Learning the language of gene expression (2007, January 19) retrieved 2 May 2024 from https://phys.org/news/2007-01-language-gene.html