MedicalXpress

Bioinformatics researchers explore additional coding potential hidden in the human genome

June 18 2012

Sequencing the human genome was just the first step. The next challenge is of the kind that makes history: to decode the genome, and understand how the information needed to construct a human being can be packaged into a single molecule. And there are a lot more than loose ends in the way of a solution. A group of bioinformatics experts at the Spanish National Cancer Research Centre (CNIO) in Madrid have published findings which point to still unexplored coding potential within the genome.

The substance responsible is chimeric RNA, formed not from one gene but from fragments of several. "There is growing evidence, some of it very recent, that genome coding is more complicated than we thought, and that some RNAs may combine information from two distinct genes," explains Alfonso Valencia, head of the CNIO's Structural Biology and Biocomputing Program. "We have called them chimeric RNAs after the mythological beings made from the parts of two different animals," he relates.

The research has been carried out in collaboration with scientists from the Centro de Regulación Genómica (CRG) in Barcelona. "We noted the prevalence of this phenomenon back in 2006, and are now working to establish its biological importance", remarks Roderic Guigó, coordinator of the CRG's Bioinformatics and Genomics program.

DNA contains the genes, which are translated into proteins. RNA, meantime, serves as an intermediary molecule performing what is an indispensable step in the process: before a gene can be translated into a protein, the right RNA has to be built. The classical vision of how information is stored in the genome holds that the correspondence is one-to-one, that is: one gene, one RNA, one protein.

A PARADIGM SHIFT NEEDED

And that was what scientists expected to find when they sequenced the genome at the start of the last decade. But it was quickly apparent that there was a problem: the human genome contains some 20,000 genes, while the variety of proteins in the human body is considerably greater. Something was wrong.

We now know that a single gene can produce several proteins; just as a words like "bat", "foot" or "count" can have different meanings despite being written the same way. But it remains to be seen whether this is a common phenomenon – whether all genes can code for multiple proteins – or a rarity. In fact, here too Valencia's group has made advances, demonstrating in a paper published last April in Molecular Evolution Biology that the translation of a single gene into several proteins occurs, but is fairly uncommon.

Chimeric RNA is also partly responsible for there being more distinct proteins than there are genes. As if the system reading and translating the genes could find three or more meanings from any two. So, for example, "love" and "cast" would be direct translations, but we would also get "ve-st"; "ca-ve"; "lo-st"...

The existence of chimeric RNA was already an established fact, and it was also known that some chimeric RNAs are translated into proteins, while others remain in the RNA phase, as happens with normal, non-

chimeric RNA. But chimeric proteins were generally believed to be a rarity confined to pathological processes like the development of cancer.

The CNIO's bioinformatics team trawled through gene, RNA and protein databases and conducted new experiments before finally discovering that chimeric RNA is present in far greater quantities than was first thought. They have also detected cases of translation to proteins as part of an apparently normal process in healthy as well as cancerous tissue.

Their results have been written up in a series of papers, the latest of which has just appeared in *Genome Research* (Frenkel-Morgenstern et al, 2012, PMID: 22588898) signed as first author by Milana Frenkel-Morgenstern, from the CNIO Structural Computational Biology Group that Valencia leads. The interest has been such that another journal, *Nature Reviews Genetics*, dedicated a commentary to the article (Post transcriptional regulation: Chimeric protein production, NRG, June 7, 2012, 10.1038/nrg3268).

Specifically, the CNIO researchers have identified 175 chimeric RNA transcripts in 16 human tissues, and 12 new chimeric proteins. This finding poses numerous questions: How important is this process out of all the information in the genome? Does it finally explain the mismatch between the number of genes and proteins? What is the total number of chimeric proteins? Is there some function that characterises them? Why do they exist?

"We have opened up a line of inquiry which we hope other groups will now pursue," remarks Valencia. "In my opinion, the key thing about this research is that it shows we still have a lot to learn before we fully understand what is written in the genome."