PHYS ORG

Scientists unveil new tool to understand evolution of multi-domain genes

May 16 2008

Carnegie Mellon scientists have discovered critical flaws in the standard method used to analyze gene evolution. Standard methods fail when applied to genes that encode multi-domain proteins, an important class of proteins crucial to human health.

Computational biologist Dannie Durand and colleagues have for the first time tackled the dilemma of how to study the ancestry of multi-domain genes.

Correctly identifying gene ancestry is a linchpin of computational genomics. Genes passed down from a common ancestor tend to perform similar functions in the cell. Scientists exploit this similarity to perform tasks such as predicting gene function, mapping human chromosomal regions to corresponding regions in model organisms, and reconstructing the regulatory circuitry that turns genes on and off.

Although computational biologists have developed methods to identify genes that share a common ancestor, current methods often lead to spurious conclusions when applied genes encode multi-domain proteins. Domains are sequence fragments that encode the basic building blocks of protein structure. Evolution makes new genes by mixing and matching domains in novel combinations, much like a child who builds a house, a car and a helicopter from the same LEGO kit by combining LEGO blocks in different ways. This process, called domain shuffling, creates complex proteins that perform specific, critical tasks such as cell communication and binding to other cells. When one of these proteins fails, cancer is often the result. Domain shuffling allows rapid evolution of new proteins, but it also makes it close to impossible for scientists to determine their ancestry.

In a paper published online in Public Library of Science *Computational Biology* today, Durand's team presents a novel method to determine whether a pair of similar genes evolved from a common ancestor, or whether they just look similar because the same domain was inserted into both genes. Their method, called "Neighborhood Correlation," is the first to tackle this problem.

"We needed a completely new approach to determine which multi-domain proteins share a common ancestor, and we are the first group to propose such a method," Durand said. "Ours is the first approach to define and analyze common ancestry in a traditional vertical way, even when domain shuffling occurs."

Neighborhood Correlation exploits the structure of a statistically weighted sequence similarity network to differentiate multi-domain genes with shared ancestries from multi-domain genes that result from domain shuffling. Gene duplication creates a specific signature in the network, while domain insertion creates a different characteristic signature. Neighborhood Correlation captures these signatures, giving pairs that arose through duplication, and hence share common ancestry, a higher score than genes that share an inserted domain, but not a common ancestor.

The Carnegie Mellon scientists tested Neighborhood Correlation against 20 protein families — including Kinases, the largest multi-domain family found in humans — whose ancestral relationships are well established through lab-based research. The tool worked remarkably well in verifying the ancestral patterns of multi-domain gene evolution for these families, much better than the tools we use today, Durand said.

Today’s computational tools use sequence similarity, assuming that genes with similar sequences indicate common ancestry. Those methods also use the length of the similar region to rule out similarity that arose due to inserted domains. They reason that the longer the sequence shared by two multi-domain genes, the more likely that those two genes share a common ancestor.

But Durand’s tests showed that this assumption often does not hold. Her team found disturbing results when they compared sequence similarity to their Neighborhood Correlation method in evaluating the 20 gene families with established histories. The sequence similarity method actually yielded false ancestral associations and missed true ancestral relationships.

Neighborhood Correlation is successful because it takes both gene duplication and domain insertion into account.

“Not only do we show that Neighborhood Correlation works empirically, we also provide a sound evolutionary argument as to why it should work,” Durand observed. “Our results show that the organization of sequence similarity network contains evidence of ancient evolutionary processes. This has exciting implications for future studies. We hope that comparing the sequence similarity networks of different species will reveal how evolutionary processes differ in plants, animals and fungi,” Durand said. “Multicellularity evolved independently in each of those groups. To go from a single cell to many cells acting together, each time nature had to solve the same problems of cellular communication and control. But are the solutions the same in each lineage" How those problems were solved is a fascinating question.”

Although designed for multi-domain families, Durand notes that Neighborhood Correlation also accurately predicts ancestry in single domain sequences. The researchers hope that scientists will begin to apply the analysis to genomic studies to better understand the role multi-domain proteins play in important evolutionary events, such as the emergence of multicellular animals and the vertebrate immune system.

Source: Carnegie Mellon University

Citation: Scientists unveil new tool to understand evolution of multi-domain genes (2008, May 16) retrieved 2 May 2024 from https://phys.org/news/2008-05-scientists-unveil-tool-evolution-multi-domain.html