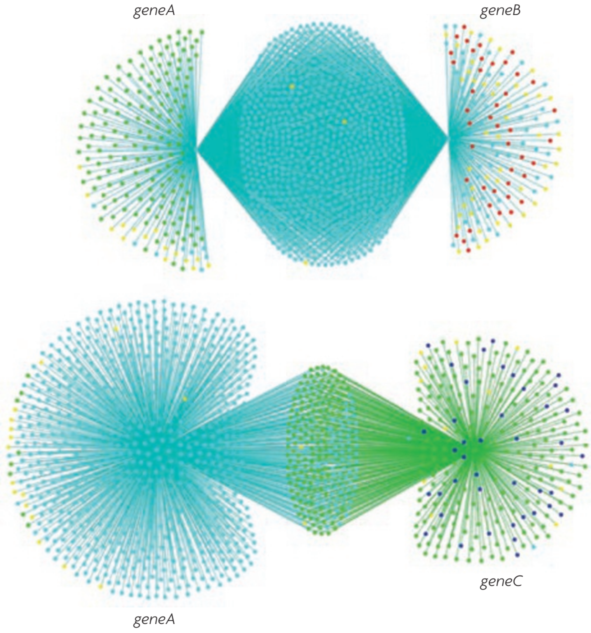
Nature Reviews Genetics | AOP, published online 10 June 2008; doi:10.1038/nrg2404

COMPUTATIONAL BIOLOGY

Assessing true homology

Understanding the ancestral relationship between genes is essential for gene annotation, prediction of function and comparative genomics. A new study questions the accuracy of current approaches for inferring gene homology, which are especially error prone when applied to complex, multidomain genes, and describes a new and robust computational method that addresses these limitations.



Differences in neighbourhood structure of the sequence similarity network reflect differences in evolutionary history. Top, a homologous match (many matches in common); bottom, a domain-only match (few matches in common). Nodes represent sequences; edges connect pairs with significant sequence similarity. Image reproduced from Song, N. *et al.* (2008).

Identifying homologous relationships between genes is typically carried out by assessing the degree of similarity between sequences. If genes contain multiple domains then inferring their evolutionary relationships is more problematic, as it is not obvious whether the genes are related by descent or whether they simply share an inserted domain — different parts of a gene might well have a distinct history. Given that a large number of proteins that are important for human health are encoded by multidomain genes, reliable methods for tracing their evolution are in great demand.

Durand and colleagues developed a homology assessment method that uses the structure of the sequence similarity network (see figure) to differentiate between homology and domain-only matches — a problem that was previously thought to be insurmountable. Their approach, called 'Neighborhood Correlation', relies on the different pattern left by homologous genes and by genes that are related by domain shuffling in a sequence similarity network.

The method was evaluated against a test data set of 20 gene families of known ancestry from humans and mice, giving a total of >850,000 pairwise interactions. Not only did the newly developed program perform well (for example, it correctly classified 12 families more accurately than any other method), it also revealed that existing homology assessment methods were error prone: these methods often inferred incorrect evolutionary relationships or missed genuine ones. Even more strikingly, they also found that programs that were deliberately designed to infer the ancestry of multidomain proteins fared worse than more general methods.

Neighborhood Correlation is the first approach to assess explicitly whether multidomain genes have a common ancestry. Although the method was specifically designed to help with multidomain genes, it is just as effective when used on single-domain sequences. Given the high degree of imprecision of current methods that was highlighted in the course of this work, the authors suggest that it might be appropriate to re-evaluate the conclusions of some existing studies.

Tanita Casci

ORIGINAL RESEARCH PAPER Song, N. *et al.* Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comp. Biol.* **4**, e1000063 (2008)
WEB SITE
Neighborhood Correlation:
http://www.neighborhoodcorrelation.org