Detection of gene and sub-gene differential conservation patterns in complete proteomes

Audrey Defosset[1], Arnaud Kress[1], Yannis Nevers[1], Raymond Ripp[1], Olivier Poch[1], Odile Lecompte[1]

Context

In recent years, advances in sequencing technologies have led to a vast increase in genomic data, as well as a wider diversity of available species, providing opportunities to better study the evolution of various biological processes, and understand genotype/phenotype relations. Comparative genomics has been established as a particularly well suited approach to exploit genome sequencing data, and especially phylogenetic profiling, which is commonly used to compare gene repertoires between several species. It is admitted that genes participating in the same mechanism will generally be conserved and lost together through evolution, and that functionally linked genes often present similar taxonomic distributions (Pellegrini et al., 1999). It is thus possible to infer gene function and associate genes to various processes by matching a phenotype distribution to that of a set of genes.

While phylogenetic profiling is a very robust approach to explore evolutionary histories of species, it does not account for the modular nature of protein domain evolution that can be observed. It has indeed been shown that domain gains and losses are quite common through evolution, and multi-domain protein architecture is often rearranged between taxa, participating in lineage specific adaptations (Moore and Bornberg-Bauer, 2012; Zmasek and Godzik, 2011). Domain or motif variation is also true when dealing with proteins involved in multiple processes, such as moonlighting proteins, which can exhibit several biological functions (Jeffery, 1999). This makes it clear that, while it is important to consider gain and loss of complete genes, it is also crucial to take into account the domain composition and divergences in orthologs to gain better insight into the complex relations between phenotype and genotype, and potentially predict specializations and phenotype divergences between related species.

A few tools have been developed with the aim of integrating domain variations to evolutionary studies, however, most are either limited to the study of individual genes or genes families, and don't seem to be adapted to the exploitation of the massive amount of data currently available. These programs also mostly focus on well characterized functional domains such as PFAM domains, which prevents the analysis of uncharacterized domains or small motifs and don't allow the detection of subtle sequence divergence, which have been shown to be able to alter domain function entirely (Anderson et al., 2016).

To answer the need for a high-throughput method capable of detecting divergent conservation patterns at the sub-gene level, we developed a new approach based on BLAST homology searches (Altschul et al., 1997) to identify, in a whole proteome, proteins presenting lineage-specific genotype divergences resulting from domain or motif gain/loss or variation.

Results and discussion

We designed BLUR (Blast Unexpected Ranking), a novel approach capable of detecting both gene gain/loss and sub-gene levels divergences in two selected groups of species, with the aim of predicting phenotype divergences or specializations between lineages, or find genes involved in known phenotype differences. It is based on the assumption that in a BLAST result, the succession of hits approximately respects a defined taxonomic order, whereas for proteins presenting a missing or divergent region, that order will be altered. The BLUR program allows the analysis of two related taxa, by establishing their standard conservation behavior when compared to a more distant reference proteome using BLASTP. BLAST searches are precomputed for the whole proteome of the reference species (query proteome). The global rankings in the BLAST results are then compared for both taxa, as well as E-values and distances to the query, followed by a statistical analysis to determine cases where a divergence would be atypical.

Our approach was successfully tested on several biological questions, including the detection of cilia-related genes in Eukaryotes by comparing genes of ciliated and non-ciliated Fungi. Using BLAST searches performed on the human proteome, we compared non-ciliated Dikarya to six ciliated Fungi, highlighting a set of 1179 proteins highly enriched in ciliated GO terms, among which 1081 are absent in Dikarya and 98 are divergent when compared to ciliated Fungi.

We developed a website (www.lbgi.fr/blur/) to make the use of BLUR easy and straightforward. 27 reference proteomes are available and over 4000 species from all three life domains can be compared to one another. The website provides the opportunity of both global and individual analyses of the results, such as the generation of interaction networks using STRING data, or Gene Ontology enrichments, as well as multiple sequence alignment for each protein, allowing for a better visualization of the detected sequence divergences.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402.

Anderson, D.P., Whitney, D.S., Hanson-Smith, V., Woznica, A., Campodonico-Burnett, W., Volkman, B.F., King, N., Thornton, J.W., and Prehoda, K.E. (2016). Evolution of an ancient protein function involved in organized multicellularity in animals. ELife 5.

Jeffery, C.J. (1999). Moonlighting proteins. Trends in Biochemical Sciences 24, 8–11.

Lees, J.G., Dawson, N.L., Sillitoe, I., and Orengo, C.A. (2016). Functional innovation from changes in protein domains and their combinations. Current Opinion in Structural Biology 38, 44–52.

Moore, A.D., and Bornberg-Bauer, E. (2012). The Dynamics and Evolutionary Potential of Domain Loss and Emergence. Mol Biol Evol 29, 787–796.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A 96, 4285–4288.

Zmasek, C.M., and Godzik, A. (2011). Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol 12, R4.

1. Contact: adefosset@etu.unistra.fr, Complex Systems and Translational Bioinformatics, ICube UMR 7357, Université de Strasbourg, Strasbourg, France