Real or fake? Verifying protein domain gain and loss events by an automated fact-checking approach

Julie D. Thompson[1], Arnaud Kress[1], Olivier Poch[1], and Odile Lecompte[1]

[1]Universite de Strasbourg

February 11, 2023

Abstract

The comparison of protein domain architectures provides insight into the evolution and function of proteins. By comparing the domains of different proteins, scientists can identify common domains, classify proteins based on their domain architecture, and highlight proteins that have evolved differently in one or more species or clades. Such proteins are often thought to represent genetic novelty underlying unique adaptations. However, genome-wide identification of different protein domain architectures involves a complex error-prone pipeline that includes genome sequencing, prediction of gene exon/intron structures, and inference of protein sequences and domain annotations. Here we developed an automated fact-checking approach to distinguish true domain loss/gain events from false events caused by errors that occur during the annotation process. Using genome-wide ortholog sets and taking advantage of the high-quality human and *Saccharomyces cerevisiae* genome annotations, we analyzed the domain gain and loss events in the predicted proteomes of 9 non-human primates (NHP) and 20 non- *S. cerevisiae* fungi (NSF) as annotated in the Uniprot and Interpro databases. Our approach allowed us to quantify the impact of errors on estimates of protein domain gains and losses, and we show that domain losses are over-estimated ten-fold and three-fold in the NHP and NSF proteins respectively. This is in line with previous studies of gene-level losses, where sequencing issues or incorrect gene prediction led to genes being falsely inferred as absent. For the first time, to our knowledge, we show that domain gains are also over-estimated by three-fold and two-fold respectively in NHP and NSF proteins. Based on our more accurate estimates, we infer that true domain losses and gains in NHP with respect to humans are observed at similar rates, while domains gains in the more divergent NSF are observed twice as frequently as domain losses with respect to *S. cerevisiae*. This study highlights the need to critically examine the scientific validity of protein annotations, and represents a significant step toward scalable computational fact-checking methods that may one day mitigate the propagation of wrong information in protein databases.

Hosted file

`main_text.docx` available at https://authorea.com/users/585132/articles/623971-real-or-fake-verifying-protein-domain-gain-and-loss-events-by-an-automated-fact-checking-approach