SEQrutinator: Fast and objective sequence scrutiny for the analysis of highly complex protein superfamilies.

Agustín Amalfitano, Hugo Marcelo Atencio, Nicolás Stocchi, Arjen ten Have and Fernando Villarreal.

Background Protein superfamilies can be defined as protein families that consist of paralogous subfamilies with different functioning. Proteins active in secondary metabolism show high levels of functional diversification that led to complex superfamilies such as Cytochrome P450 (CYP), superfamilies with many paralogues with unknown functions and therefore paradigms of structure-function prediction analyses. The high sequence variation of these superfamilies furthermore hampers obtaining reliable multiple sequence alignments (MSAs). This problem is exacerbated by the presence of sequences derived from incorrect gene models and pseudogenes.
Objective Obtain an objective method that allows automated sequence mining of complex superfamilies.
Method We devised SEQrutinator that subsequently detects and removes 1 Short sequences; 2 Non-homologous hits; 3 Sequences that instigate large contiguous gaps in MSAs; 4 Sequences that have large gaps in MSAs; 5 Sequence distance outliers. SEQrutinator was applied to three major protein superfamilies involved in plant secondary metabolism: CYP, UDP-Glycosyl Transferases (UGT) and Acyl Transferases (AT). We scrutinized 16 complete proteomes, using SwissProt as a control. A recovery screening was implemented to recuperate sequences that form distant subfamilies.
Results We compared the performance of 17 separate SEQrutinator outputs. We detected a small number of incorrect, mostly partial sequences in SwissProt. Performance of SEQrutinator depended mostly on the proteome: Only few sequences were removed from the highly curated plant model *Arabidopsis thaliana* whereas the more likely poorly curated proteomes, such as Gibko and Norway spruce, showed many instances of incorrect sequences. The recovery screen, that we applied to a combined dataset of 9 complete plant proteomes identified a single distant subfamily in each superfamily. We recovered CYP74, a functional subfamily that, besides a relatively large evolutionary distance, has a small number of residues that interfere in the conserved heme binding motif. In the UGT case we recovered a subfamily that glycosylates sterols and lacks the plant specific pattern. In the AC case we identified an AC-like subfamily lacking one of two catalytic residues. HMMERCTTER analysis showed that scrutiny results in a large improvement of clustering training data and classification of target sets.
Prospects The method needs confirmation on a large number of protein superfamilies.