(a)

All TreeFam clean trees

Assignement of Pfam domains to protein sequences

Refinement of intitial domain predictions

Weighted parsimony

Dollo parsimony

Intersection of the gains pbtained by these two algorithms
(4,362 gained domains)

No leaf gains
(1,372 gained domains)

Representative descendant is in the genome of better quality
(653 gained domains)

No similarity with the sequences that do not have that domain assigned to them
(378 gained domains ~ 349 gained events)

High confidence domain gain events

Representative descendant is in the genome of better quality
(849 gained domains)

Medium confidence domain gain events

(b)

High confidence domain gain events

Characteristics of the gained domains

Position in a protein: N-, C-terminal or middle

Number of exons that code for the gained domain

Does domain gain extends one of existing exons or are new exons gained?

Origin of the gained domains

Blast search against the human proteoms

Exclusion of gene paralogues

Manual inspection of the gained domains with their origin

Extra analysis

Is adjacent to a homologue without the gained domain a gene with that domain?

Adjacent gene fusion? Manual inspection of candidates

What is the rate of gene duplication and speciation events after which domain gain has occurred?

Comparison with the expected rate

Medium confidence domain gai_n events

Characteristics of the gained domains

Position in a protein: N-, C-terminal or middle

Number of exons that code for the gained domain

Does domain gain extends one of existing exons or are new exons gained?