EMBL's European Bioinformatics Institute (EMBL-EBI)

Improved selection of canonical proteins for reference proteomes

Giuseppe Insana[1], Maria J. Martin[1], William R. Pearson[2]

(1) EMBL-EBI, Hinxton, UK (2) U. of Virginia, Charlottesville, VA USA

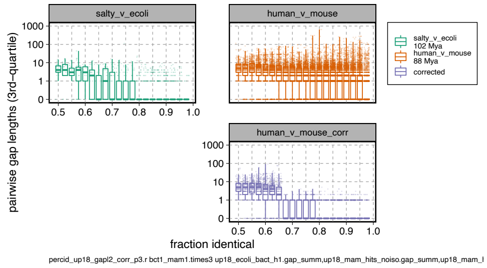
Introduction

The **Reference Proteomes** dataset seeks to provide complete proteomes for an evolutionarily diverse, less redundant, set of organisms.
As higher eukaryotes often encode multiple isoforms of a protein from a single gene, the Reference Proteomes pipeline selects a single representative ('**canonical**') sequence. UniProt identifies canonical isoforms using a *'Gene-Centric'* approach: proteins are grouped by gene-identifier and for each gene a single protein sequence is chosen.

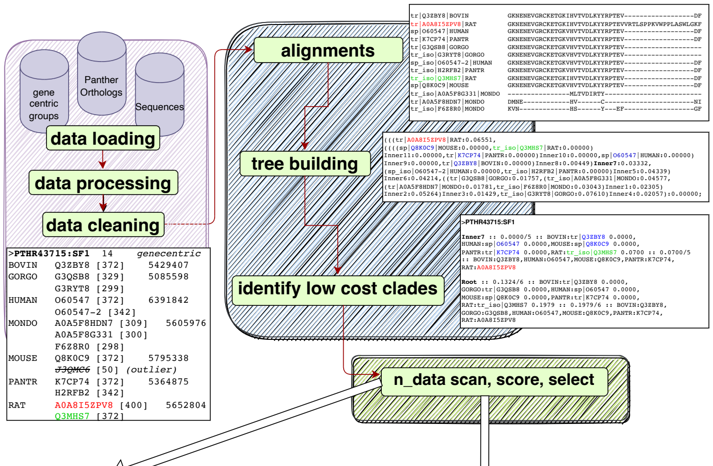
For unreviewed (UniProtKB/TrEMBL) protein sequences (and for some reviewed sequences), the longest sequence in the Gene-Centric group is chosen as canonical. This can create inconsistencies, selecting sequences with dramatically different lengths as canonical for orthologous genes. Biologically, it is unlikely that orthologous mammalian proteins differ greatly in length, but this happens about **10%** of the time for the **8 mammals in the Quest for Othologs** set.

The **Ortho2tree** data pipeline examines **Gene-Centric** canonical and isoform sequences from sets of orthologous proteins (from **PantherDB**), and suggests replacements for canonicals that have lengths very different from closely related orthologs.

Closely related (>80% id) bacterial proteins have few gaps, unlike mammalian proteins



The ortho2tree pipeline combines gene-centric isoforms of orthologs to build length-consistent canonical orthologs



```
tr|Q3ZBY8|BOVIN          GKNENEVGRCKETGKIHVTVDLKYYRPTEV-----------------DF
tr|A0A8I5ZPV8|RAT        GKNENEVGRCKETGKIHVTVDLKYYRPTEVVRTLSPPKVWPPLASWLGKF
sp|O60547|HUMAN          GKNENEVGRCKETGKVHVTVDLKYYRPTEV-----------------DF
tr|K7CP74|PANTR          GKNENEVGRCKETGKVHVTVDLKYYRPTEV-----------------DF
tr|G3QSB8|GORGO          GKNENEVGRCKETGKVHVTVDLKYYRPTEV-------------------
tr_iso|G3RYT8|GORGO      GKNENEVGRCKETGKVHVTVDLKYYRPTEV-------------------
sp_iso|O60547-2|HUMAN    GKNENEVGRCKETGKVHVTVDLKYYRPTEV-----------------DF
tr_iso|H2RFB2|PANTR      GKNENEVGRCKETGKVHVTVDLKYYRPTEV-----------------DF
tr_iso|Q3MHS7|RAT        GKNENEVGRCKETGKIHVTVDLKYYRPTEV-----------------DF
sp|Q8K0C9|MOUSE          GKNENEVGRCKETGKVHVTVDLKYYRPTEV-----------------DF
tr_iso|A0A5F8G331|MONDO  ---------------MLTVDIRTY-------------------------
tr|A0A5F8HDN7|MONDO      DMNE-----------HV-----C----------------------NI
tr_iso|F6Z8R0|MONDO      KVN------------HS-----Y---EF-----------------GF
```

```
(((tr|A0A8I5ZPV8|RAT:0.06551,
((((sp|Q8K0C9|MOUSE:0.00000,tr_iso|Q3MHS7|RAT:0.00000)
Inner11:0.00000,tr|K7CP74|PANTR:0.00000)Inner10:0.00000,sp|O60547|HUMAN:0.00000)
Inner9:0.00000,tr|Q3ZBY8|BOVIN:0.00000)Inner8:0.00449)Inner7:0.03332,
(sp_iso|O60547-2|HUMAN:0.00000,tr_iso|H2RFB2|PANTR:0.00000)Inner5:0.04339)
Inner6:0.04214,((tr|G3QSB8|GORGO:0.01757,(tr_iso|A0A5F8G331|MONDO:0.04577,
(tr|A0A5F8HDN7|MONDO:0.01781,tr_iso|F6Z8R0|MONDO:0.03043)Inner1:0.02305)
Inner2:0.05264)Inner3:0.01429,tr_iso|G3RYT8|GORGO:0.07610)Inner4:0.02057):0.00000;
```

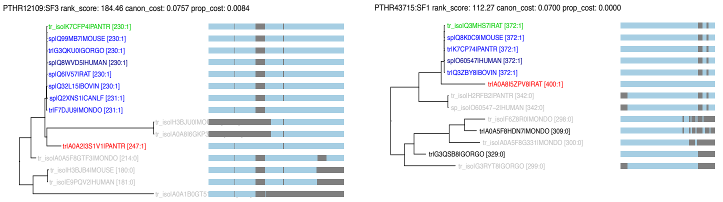
```
>PTHR43715:SF1

Inner7 :: 0.0000/5 :: BOVIN:tr|Q3ZBY8 0.0000,
HUMAN:sp|O60547 0.0000,MOUSE:sp|Q8K0C9 0.0000,
PANTR:tr|K7CP74 0.0000,RAT:tr_iso|Q3MHS7 0.0700 :: 0.0700/5
:: BOVIN:Q3ZBY8,HUMAN:O60547,MOUSE:Q8K0C9,PANTR:K7CP74,
RAT:A0A8I5ZPV8

Root :: 0.1324/6 :: BOVIN:tr|Q3ZBY8 0.0000,
GORGO:tr|G3QSB8 0.0000,HUMAN:sp|O60547 0.0000,
MOUSE:sp|Q8K0C9 0.0000,PANTR:tr|K7CP74 0.0000,
RAT:tr_iso|Q3MHS7 0.1979 :: 0.1979/6 :: BOVIN:Q3ZBY8,
GORGO:G3QSB8,HUMAN:O60547,MOUSE:Q8K0C9,PANTR:K7CP74,
RAT:A0A8I5ZPV8
```

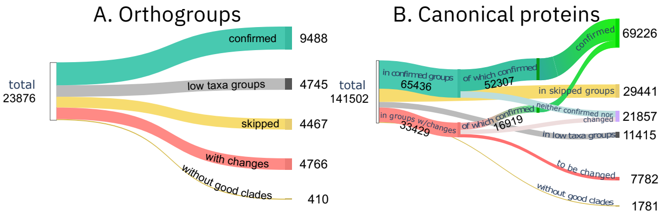
```
>PTHR43715:SF1  14      genecentric
BOVIN   Q3ZBY8 [372]     5429407
GORGO   G3QSB8 [329]     5085598
        G3RYT8 [299]
HUMAN   O60547 [372]     6391842
        O60547-2 [342]
MONDO   A0A5F8HDN7 [309]   5605976
        A0A5F8G331 [300]
        F6Z8R0 [298]
MOUSE   Q8K0C9 [372]     5795338
        ~~J3QMC6~~ [50] (outlier)
PANTR   K7CP74 [372]     5364875
        H2RFB2 [342]
RAT     A0A8I5ZPV8 [400]   5652804
        Q3MHS7 [372]
```

Gaps in vertebrate ortholog protein alignments reflect isoform choice


PTHR12109:SF3 rank_score: 184.46 canon_cost: 0.0757 prop_cost: 0.0084
tr_isolK7CFP4lPANTR [230:1]
splQ99MB7lMOUSE [230:1]
trlG3QKU0lGORGO [230:1]
splQ8WVD5lHUMAN [230:1]
splQ6lV57lRAT [230:1]
splQ32L15lBOVIN [230:1]
splQ2XNS1lCANLF [231:1]
trlF7DJU9lMONDO [231:1]
tr_isolH3BJU0lMOU
tr_isolA0A8I6GKP3
trlA0A2I3S1V1lPANTR [247:1]
tr_isolA0A5F8GTF3lMONDO [214:0]
tr_isolH3BJB4lMOUSE [180:0]
tr_isolE9PQV2lHUMAN [181:0]
tr_isolA0A1B0GT51
PTHR43715:SF1 rank_score: 112.27 canon_cost: 0.0700 prop_cost: 0.0000
tr_isolQ3MHS7lRAT [372:1]
splQ8K0C9lMOUSE [372:1]
trlK7CP74lPANTR [372:1]
splO60547lHUMAN [372:1]
trlQ3ZBY8lBOVIN [372:1]
trlA0A8I5ZPV8lRAT [400:1]
tr_isolH2RFB2lPANTR [342:0]
sp_isolO60547–2lHUMAN [342:0]
tr_isolF6Z8R0lMONDO [298:0]
trlA0A5F8HDN7lMONDO [309:0]
tr_isolA0A5F8G331lMONDO [300:0]
trlG3QSB8lGORGO [329:0]
tr_isolG3RYT8lGORGO [299:0]

Correction of chimp (A) or rat(B) canonical isoform selection using gap-based distance trees. Blue, deep-blue (MANE agreement), and red denote current canonical isoforms, red isoforms should not be canonical, they should be replaced by green isoforms.

Correction of QfO mammalian proteins



Complete results available online

fasta.bioch.virginia.edu/ortho2tree/

Methods & Results

Gap-distance trees are constructed from multiple sequence alignments of the sequences in each orthogroup. Clades with a low cost and a diverse set of taxa are identified from these trees. A weighting function is then used to seek the clades with more SwissProt entries and more canonicals, prioritising annotation from well characterized proteomes.

The pipeline suggests replacements of canonicals in the same Gene-Centric group and indicates when there is agreement between those suggestions and MANE (Matched Annotation from NCBI and EMBL-EBI) assignments.

For the ~140,000 proteins in ~24,000 Panther orthogroups from the eight mammalian proteomes (human, chimp, gorilla, mouse, rat, dog, cow and opossum), ortho2tree proposed 7782 canonical changes, while confirming 69,226 canonical assignments. When Ortho2tree suggested changing a HUMAN canonical with a MANE assignment, MANE suggested the same change 80% of the time; when no change to the HUMAN canonical was proposed, Ortho2tree and MANE agreement was 95%.

Ortho2tree can reduce canonical assignment errors among orthologous sequences that are more than 50% identical, such as sequences from vertebrates or higher plants.

National Science Foundation WHERE DISCOVERIES BEGIN

University of Virginia School of Medicine

EMBL's European Bioinformatics Institute (EMBL-EBI)
Wellcome Genome Campus, Hinxton,
Cambridgeshire. CB10 1SD. UK.
T +44(0)1223 494 444