COMPARATIVE GENOMICS AND STRUCTURAL PROTEOMICS: SEQUENCE AND STRUCTURAL HOMOLOGY TO UNRAVEL BIOLOGICAL HIDDEN FEATURES

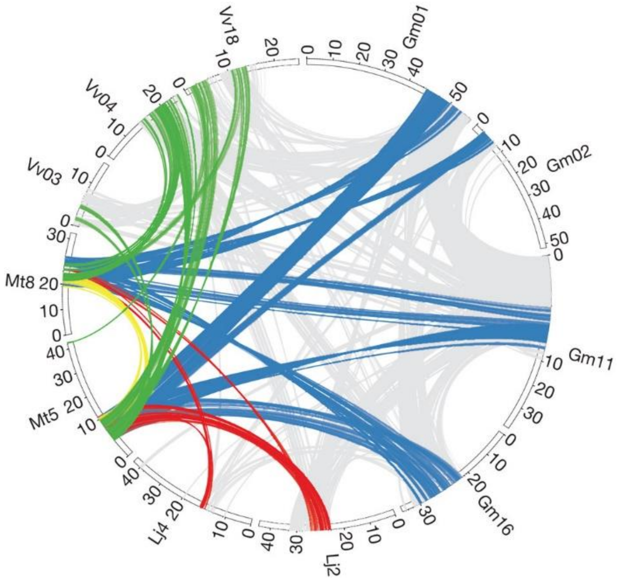
Luca Ambrosino, PhD

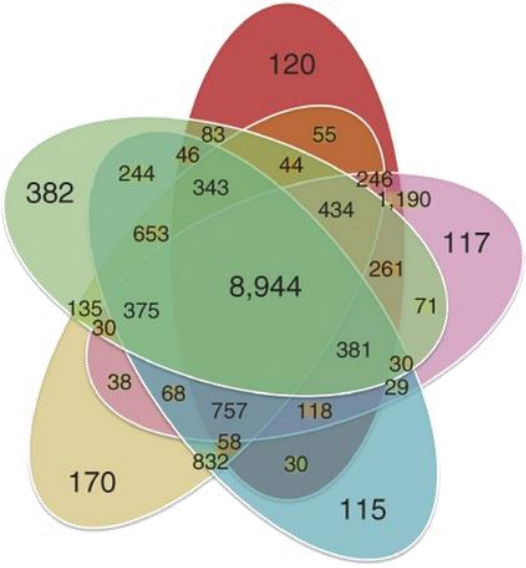
Topics

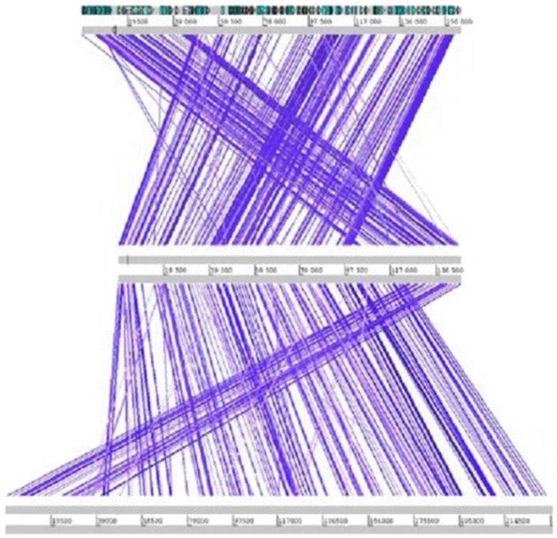
- Comparative genomics: orthologs/paralogs prediction
- Comparative genomics: functional annotation
- Structural proteomics: 3D-modeling of protein structures

Comparative Genomics

Comparative genomics studies the differences and similarities in genomic features of different organisms

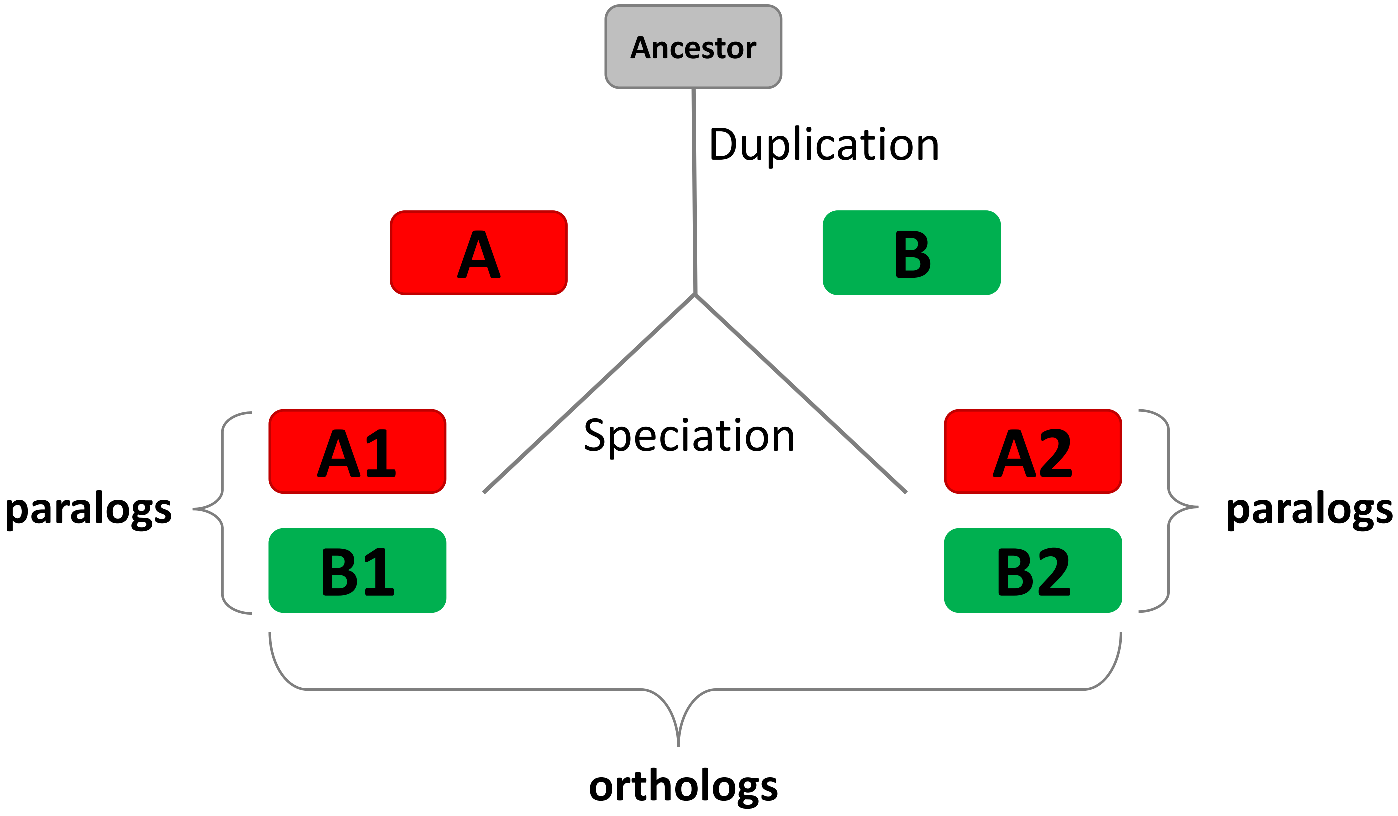




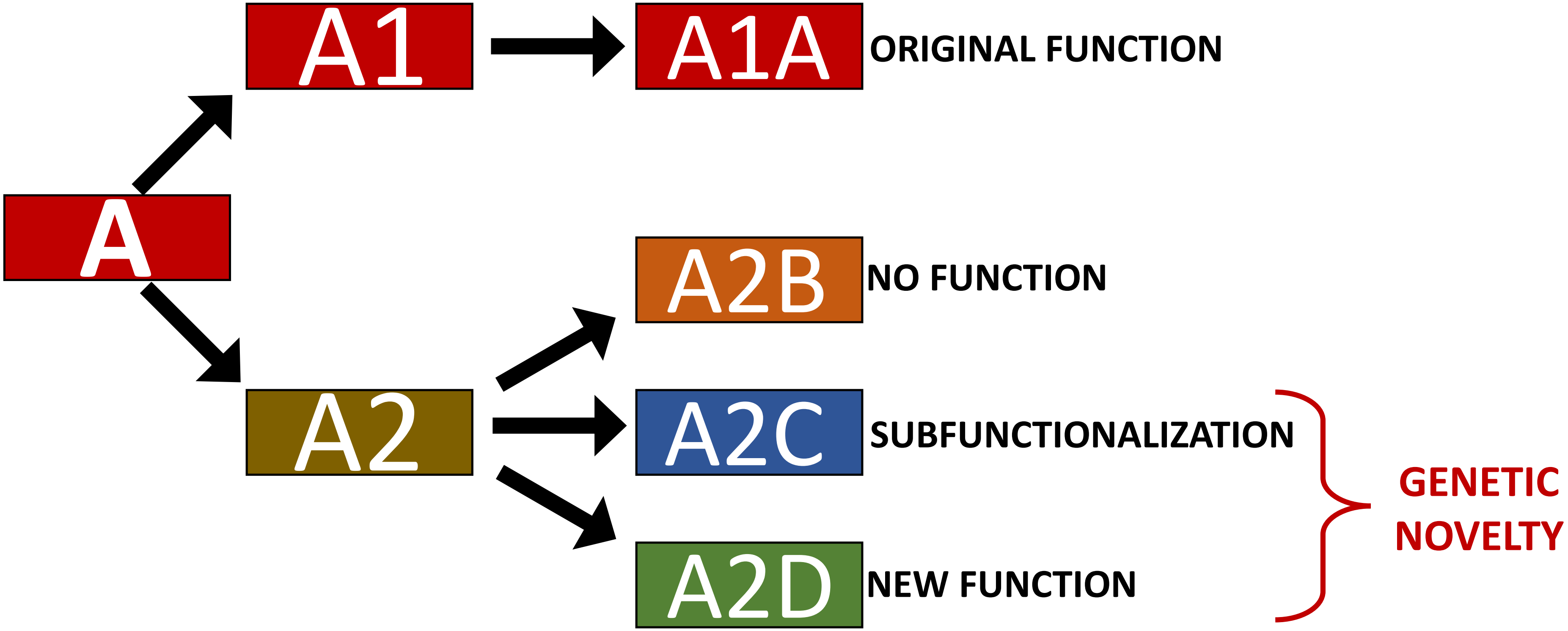


Comparative Genomics

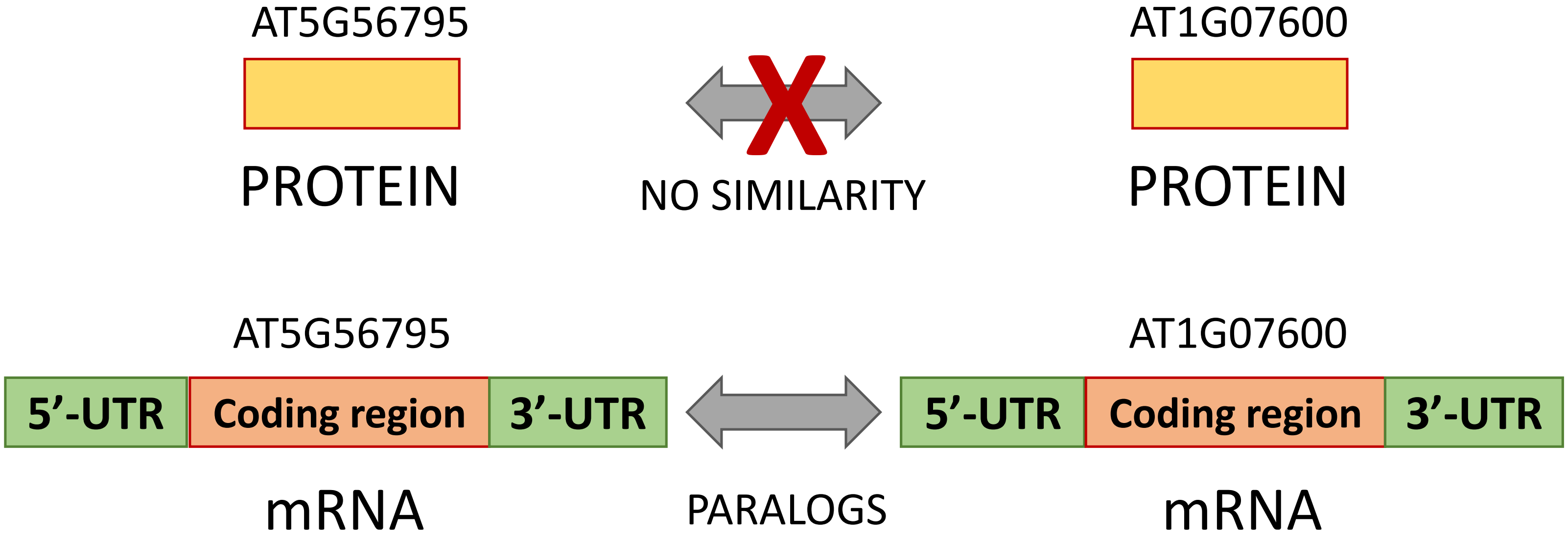
The detection of ortholog genes is one of the key approaches



Gene duplication



Multilevel analysis



AT5G56795

PROTEIN

NO SIMILARITY

AT1G07600

PROTEIN

AT5G56795

5'-UTR	Coding region	3'-UTR

mRNA

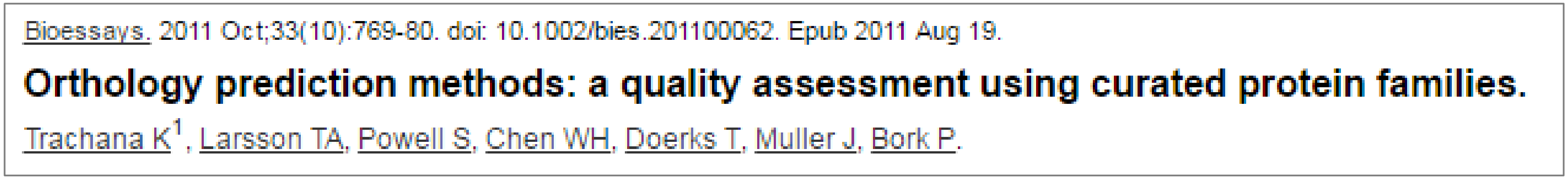
PARALOGS

AT1G07600

5'-UTR	Coding region	3'-UTR

mRNA

ORF misassignment of AT5G56795 gene

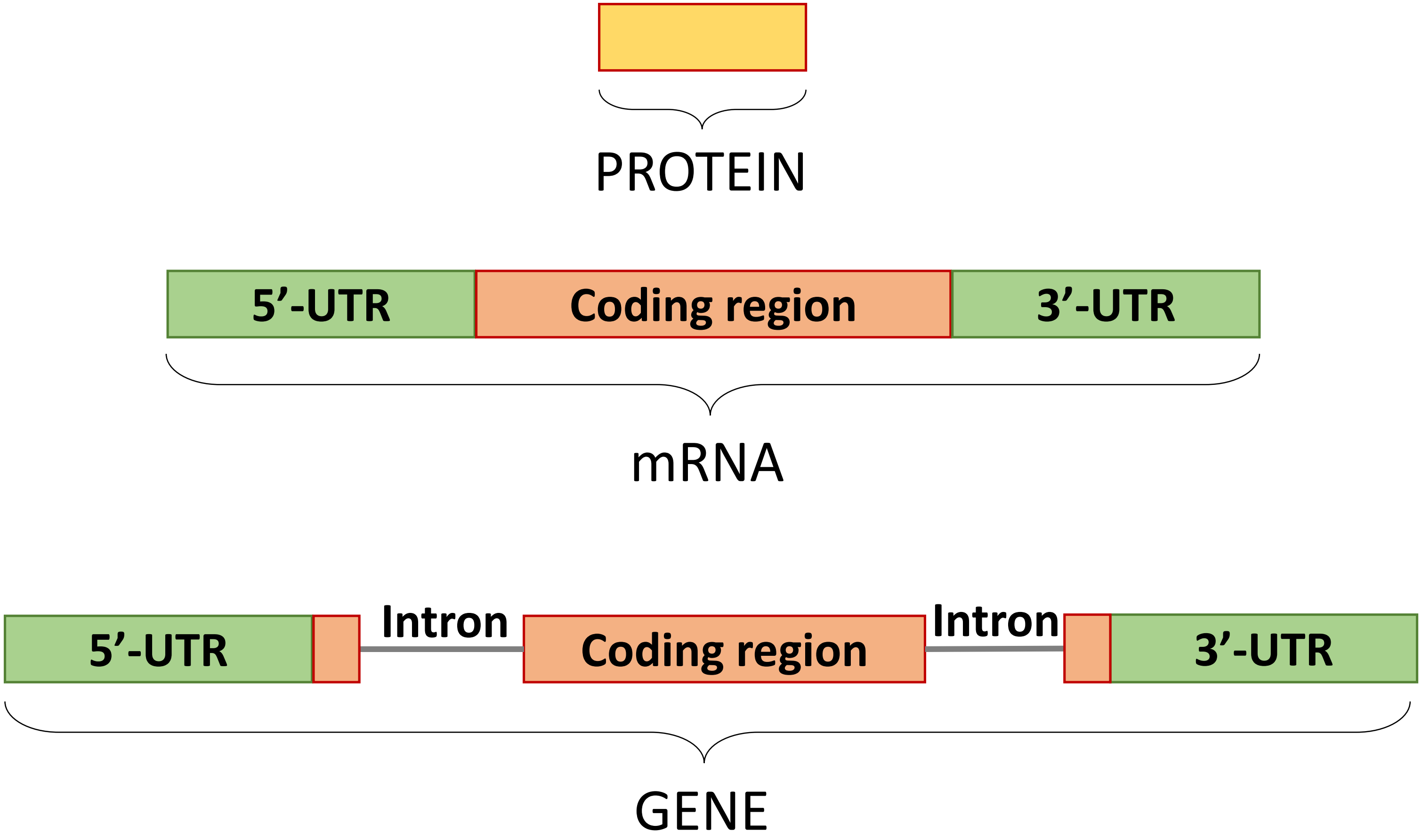


Bioessays. 2011 Oct;33(10):769-80. doi: 10.1002/bies.201100062. Epub 2011 Aug 19.

Orthology prediction methods: a quality assessment using curated protein families.

Trachana K[1], Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P.

Multilevel analysis



PROTEIN

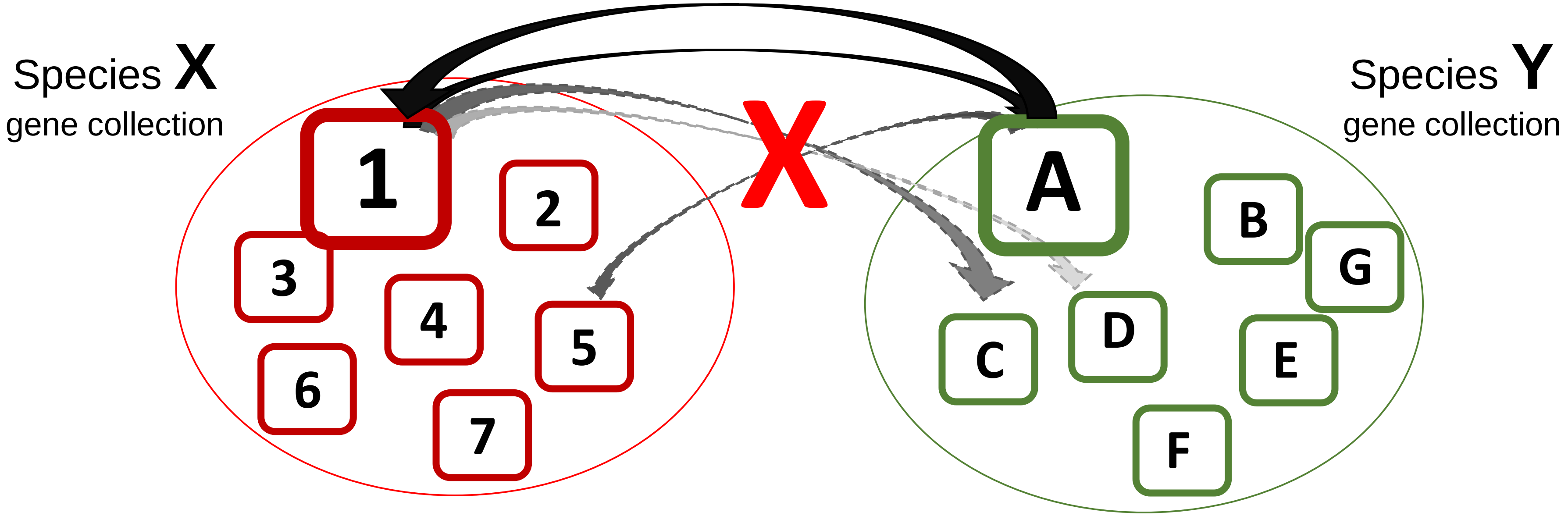
5'-UTR Coding region 3'-UTR

mRNA

5'-UTR Intron Coding region Intron 3'-UTR

GENE

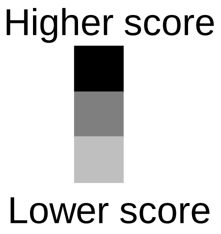
Orthologs definition



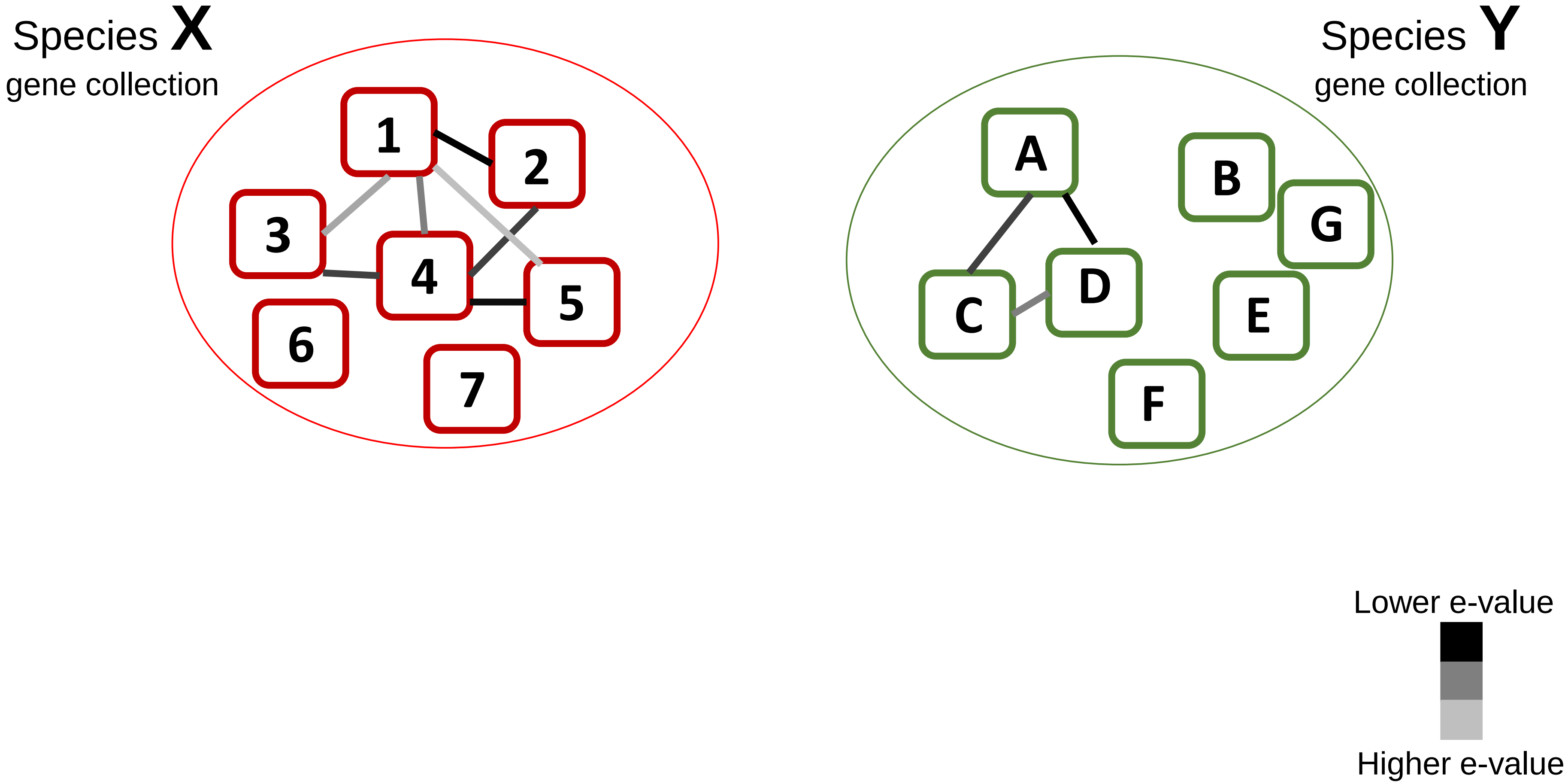

Bioinform Biol Insights. 2017 Feb 23;11:1177932217690136. doi: 10.1177/1177932217690136. eCollection 2017.

Transcriptologs: A Transcriptome-Based Approach to Predict Orthology Relationships.

Ambrosino L[1], Chiusano ML[1,2].



Paralogs definition



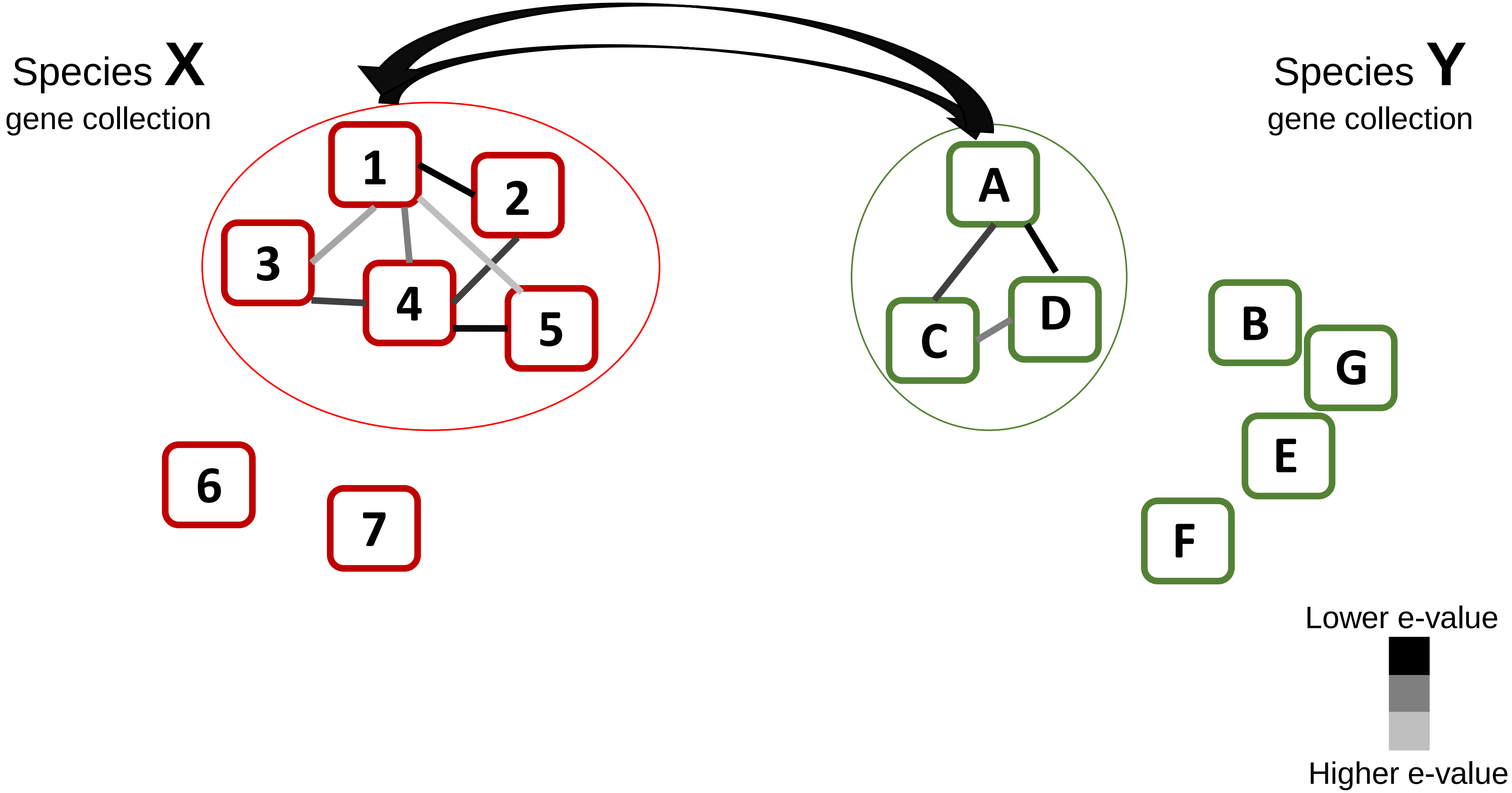
Species **X**
gene collection

Species **Y**
gene collection

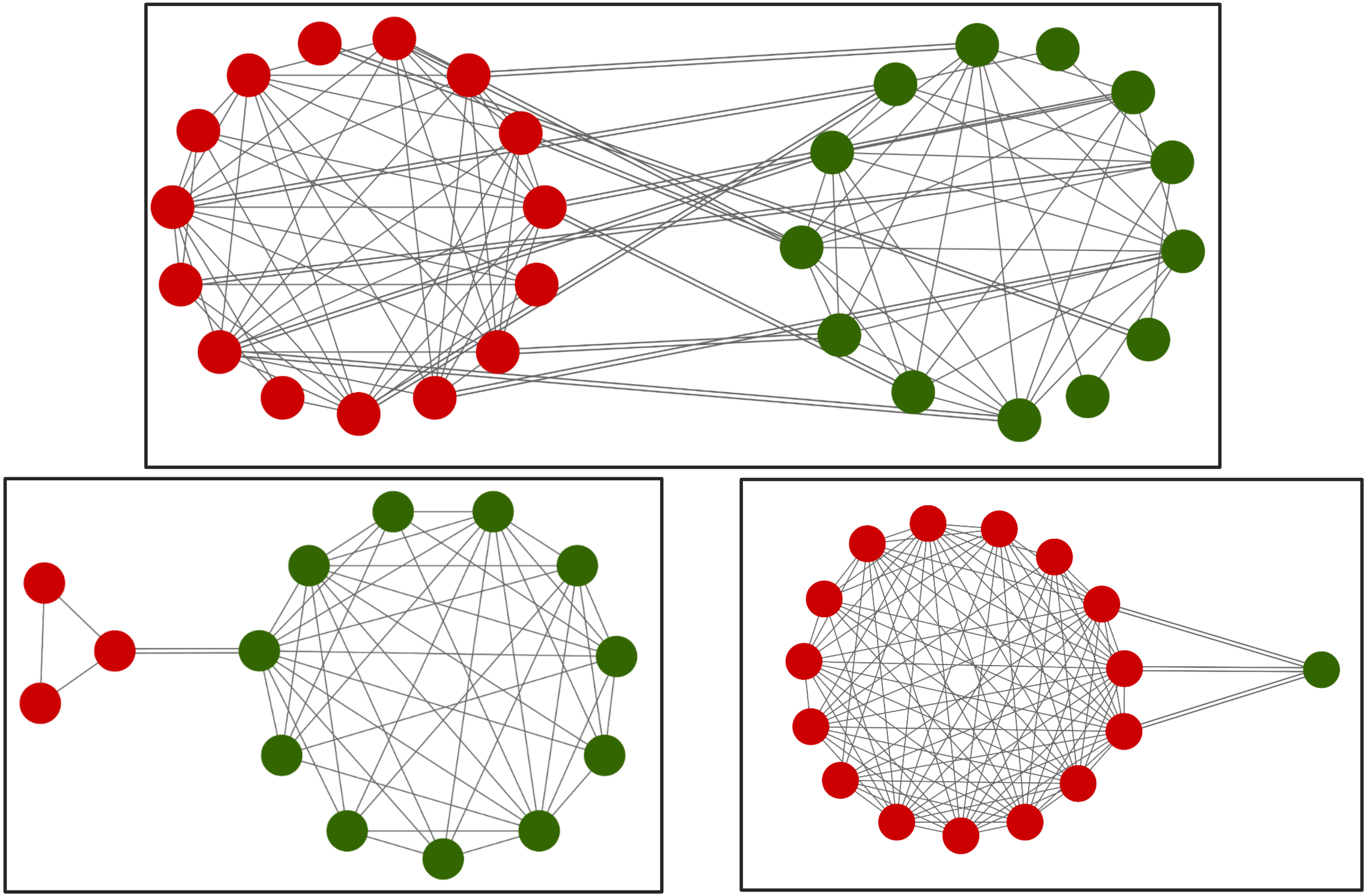
Lower e-value

Higher e-value

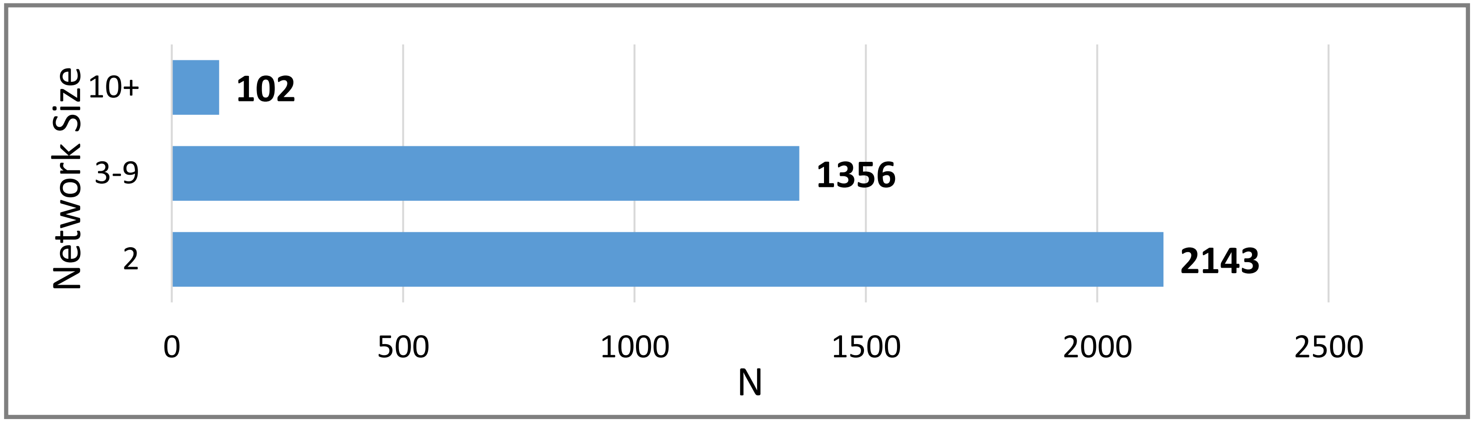
Networks construction

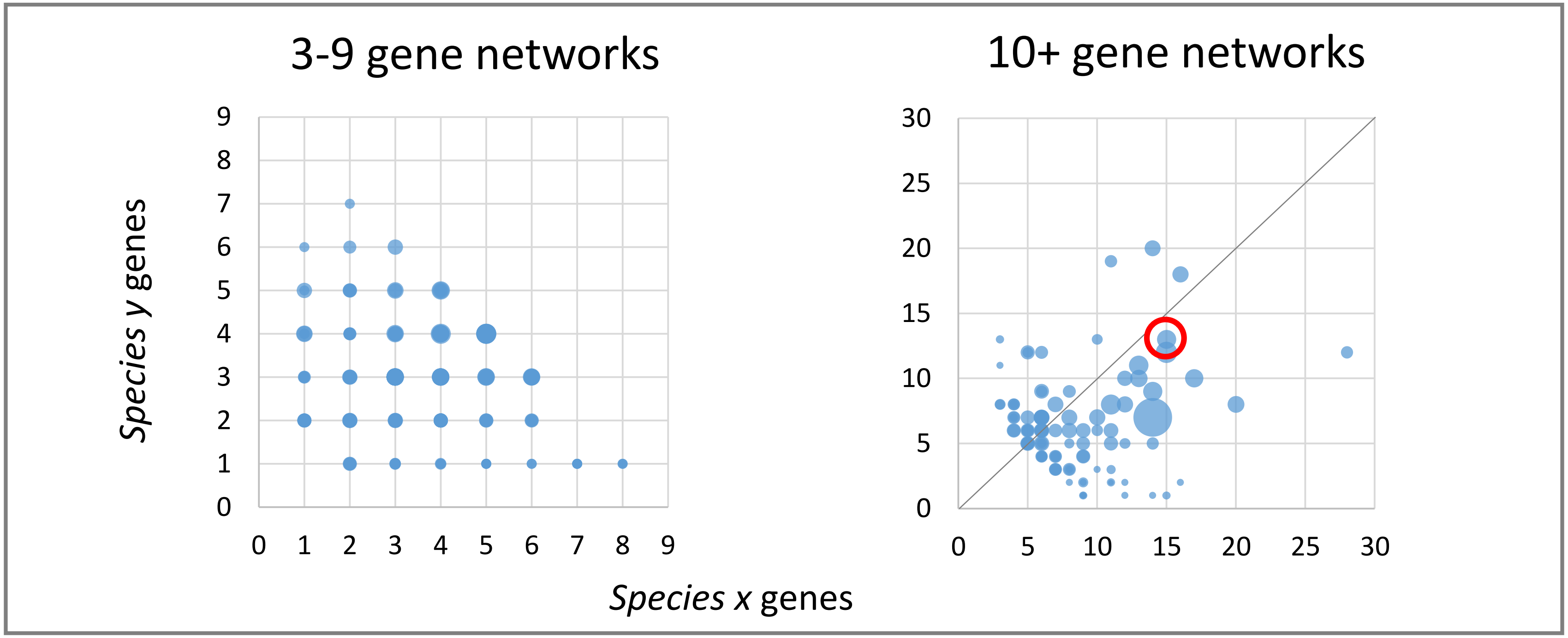


Networks construction

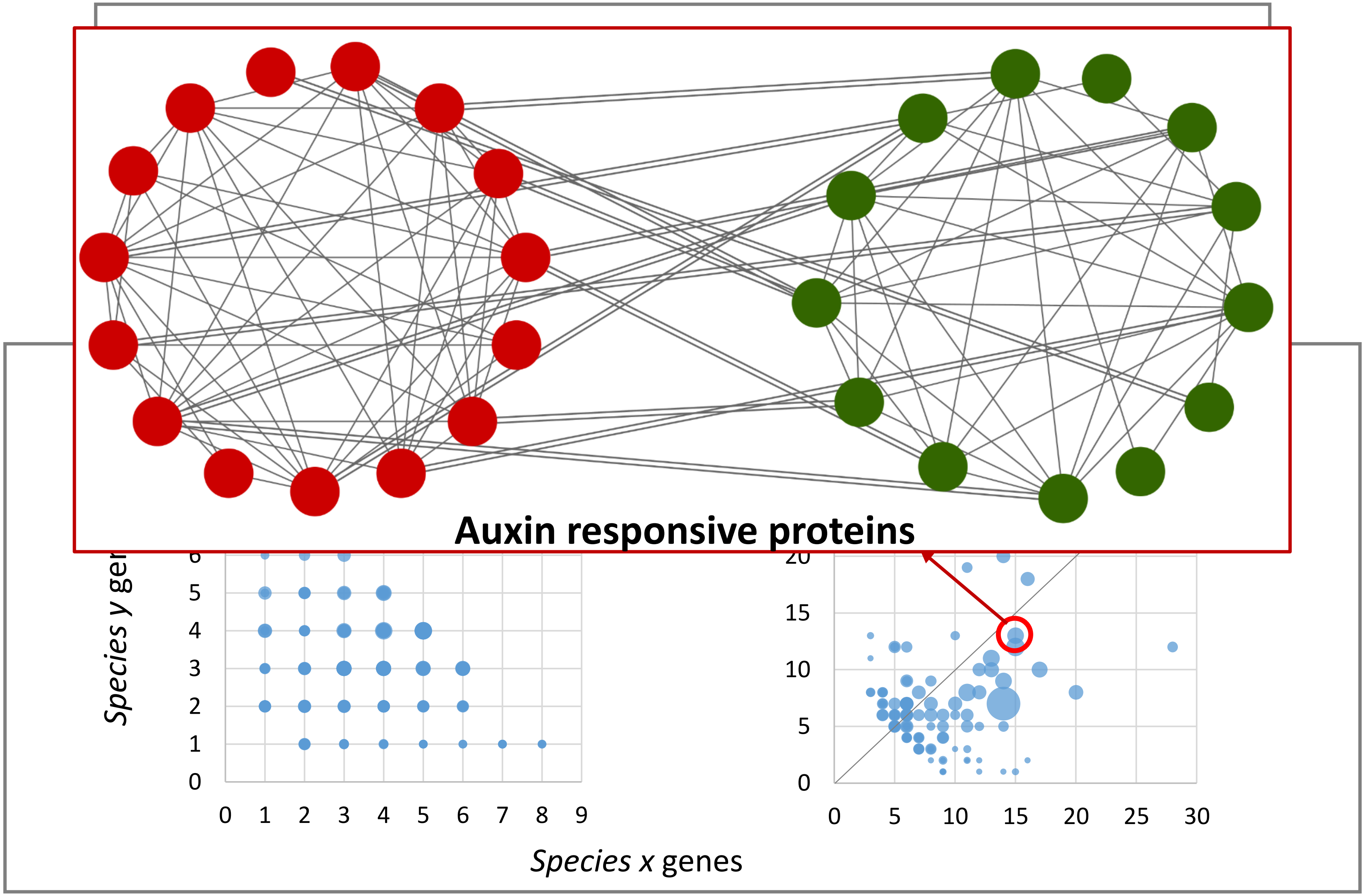


Paralog/ortholog networks

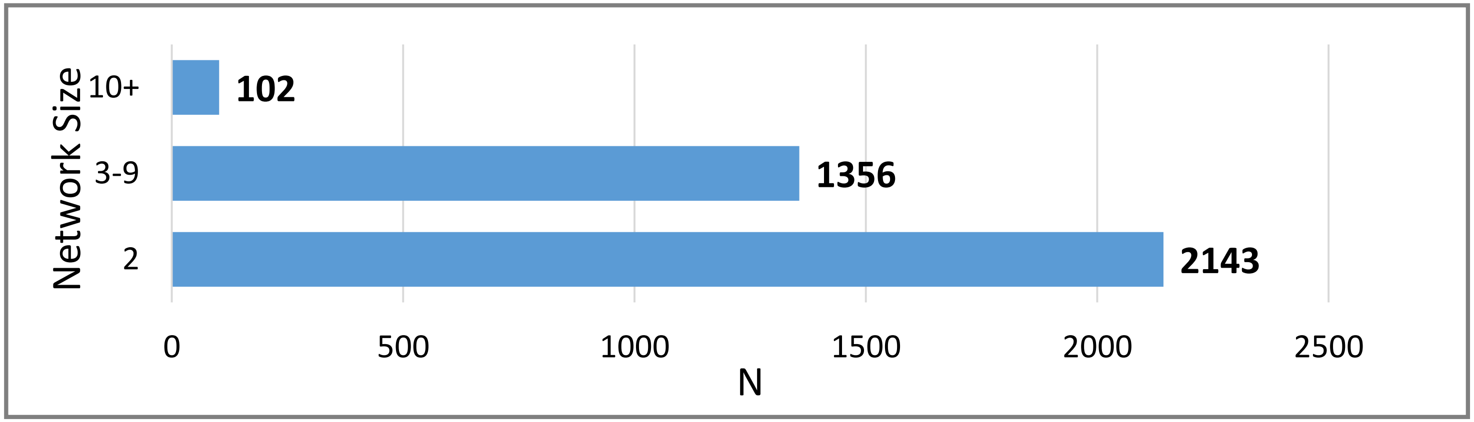


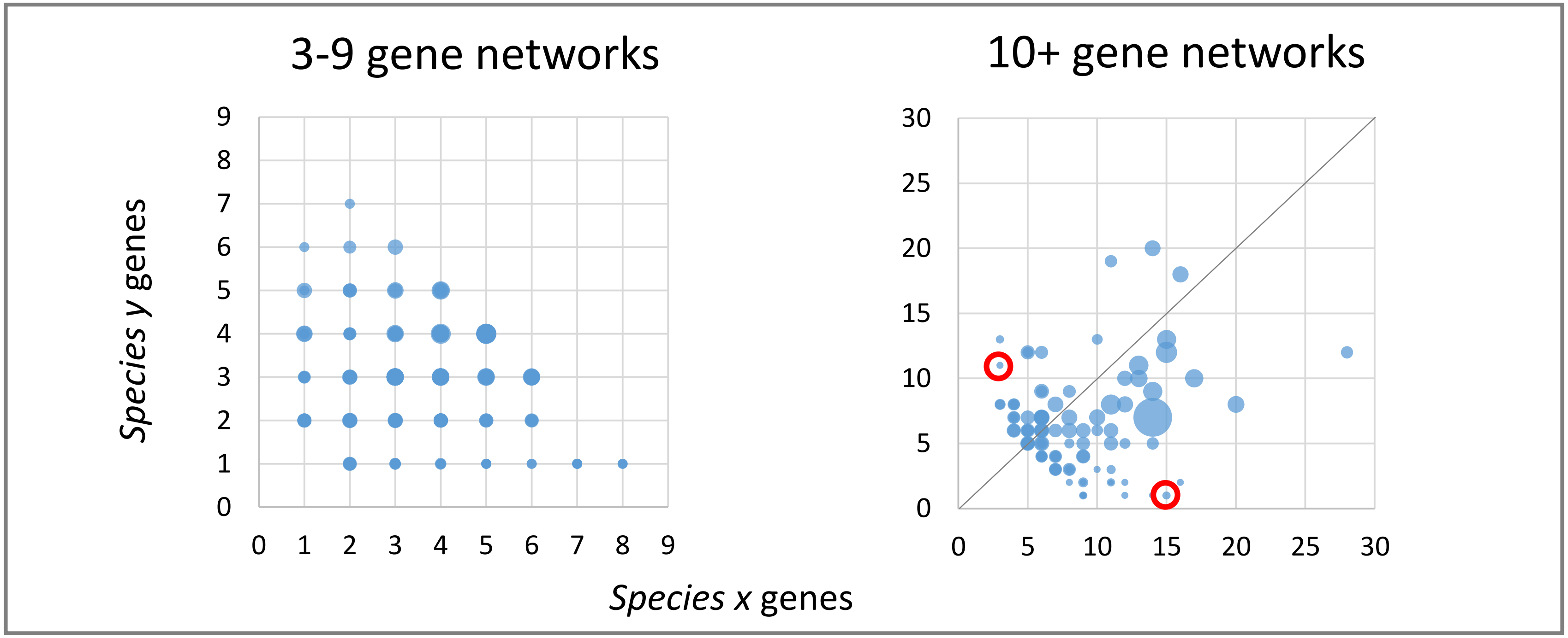


Paralog/ortholog networks

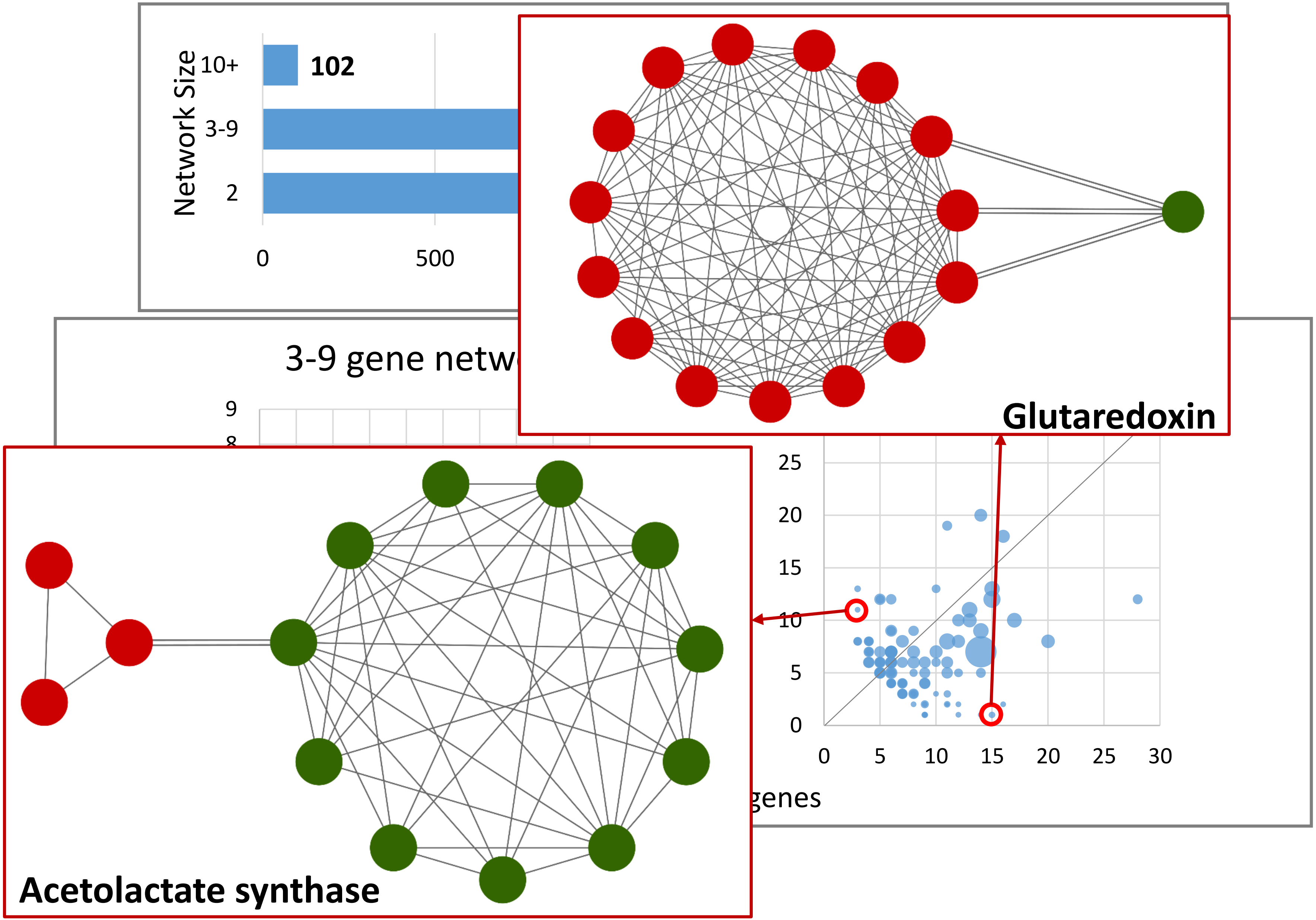


Paralog/ortholog networks

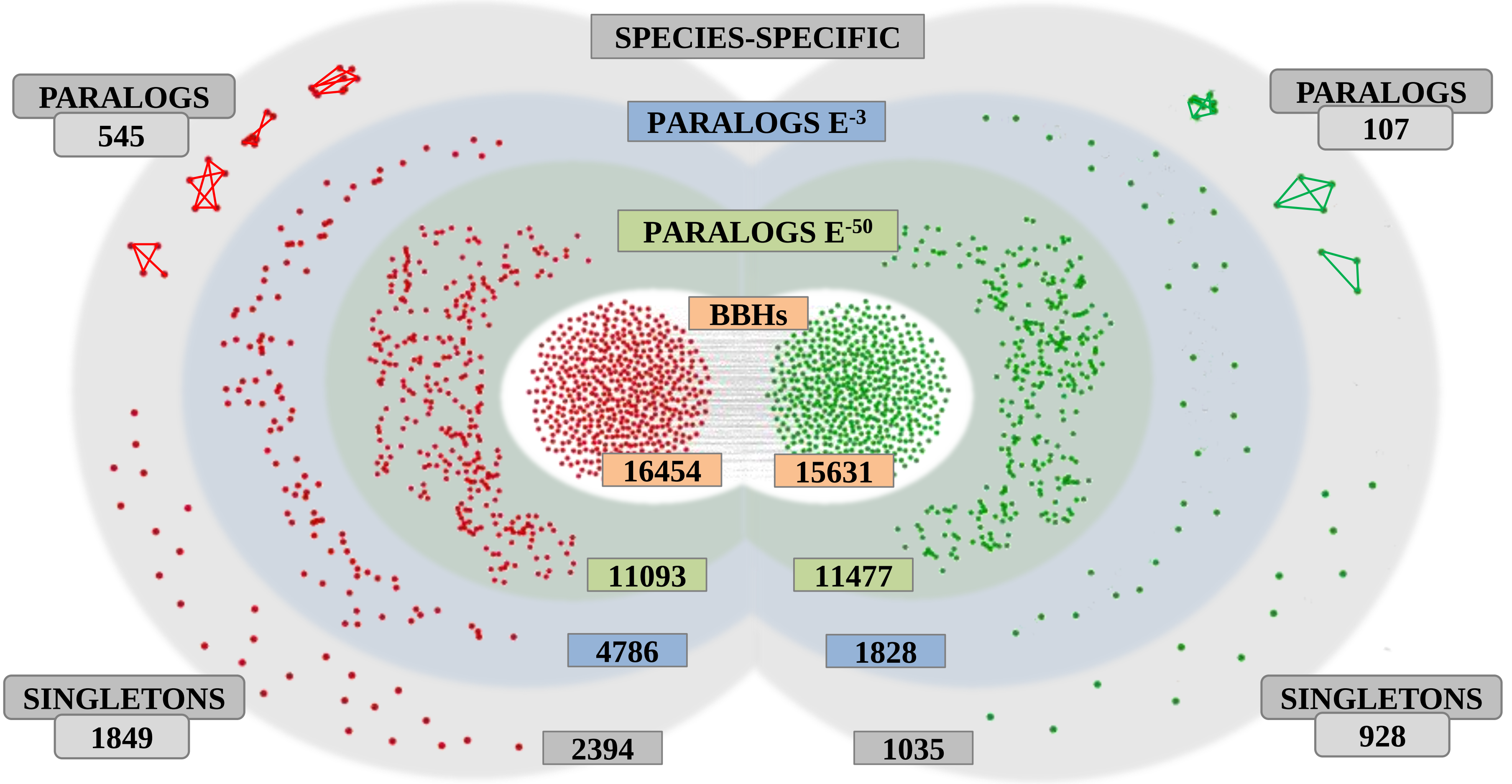




Paralog/ortholog networks



Multilevel comparison



SPECIES-SPECIFIC

PARALOGS
545

PARALOGS E^{-3}

PARALOGS
107

PARALOGS E^{-50}

BBHs

16454

15631

11093

11477

4786

1828

SINGLETONS
1849

SINGLETONS
928

2394

1035

● Species x

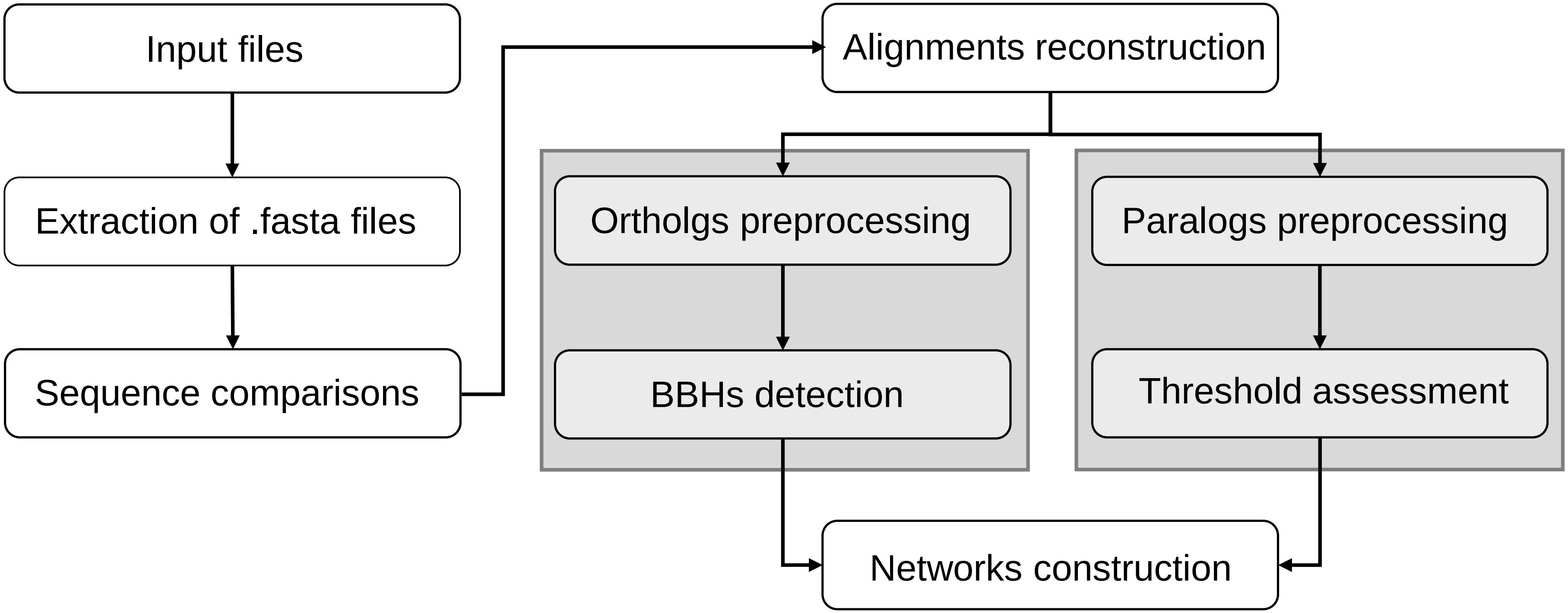
● Species y

COMPARO

COMparative for **PAR**alog and **O**rtolog genes

- Orthologs
- Paralogs
- Networks of orthologs and paralogs
- Multilevel approach

COMPARO



Input files

Extraction of .fasta files

Sequence comparisons

Alignments reconstruction

Ortholgs preprocessing

BBHs detection

Paralogs preprocessing

Threshold assessment

Networks construction

BIOINforMAtics

COMPARO

Input files

```
>NC_020166.2 Ciona intestinalis chromosome 1, KH, whole genome shotgun sequence
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAACCCCTAACTGCTAGCCCCTGCCCCcgta
accaccaacctctaacactttccaccagcccaatctgtaacgtaccaagggattcttcactagtgccacAGTTTCTAATa
cagaataaacaaaattcacTTAACTTGACCAAACAATATTCTGCTTTCAACATTATAGAACAACCTACTcactataaaaa
aacacatggcgcctgaaaaatatgtttataaaaacaatatacttATTTCGTatcaaaaacattcaaaacatATCAATTCA
AAACGCAGGCTTTACTAAACTCATTACAAAATGCGGCAAATTCAAAGTGAAGAGGTATAGTACCTCGTCTAATCGTTTTA
ATACCTAAAACTTACATACGAAATGAGTATTACCACTTAGCCACTCAAAGCAAATTGAAATGCGCGCGGCTCGAACCTCA
ATGGGGTCACGTGACAGGTCGCGGTCGGCCGATCACGAACAACAGCTGTCGAGAATCGTTTACTACGAAGCTAAGGGACG
TGTCTACTCTGTAGCATTGGGCAGATGCCGTTAGGAAAGTAACCCTTTCGCAGCAGTGTGTTTGCTGCCAAGAATAATTT
CCTCTGGCTAACAGGTGCTTTTACAGGAAGCCTGATAGGATAATTGGACACTATTTCtacattttattgttacattGTCC
AGCTTCGCTTTTTGGCACATTGGttacaattaaataaagtgttaaatatttgaGACCGTAGTTTATACACAATTCACGGC
CAGAGTTCATGCcaattttgtatatttaatgttcaaaacaaaactaaagtaTATTTCTCTCTTGCATTTATCACCGTCGA
actcaatatatagtagggtggggaagatggtacaccttaagcacataataaccaaatattctgatcgtgttttaaacaa
ttaacaacggtctat                                                  tatcaaattggaaga
gaaaatacattgaaaa                                                 cattctatttctc
gtcccatttagttgta                                                 gactgttgttaattg
tttaaaatacgatcag                                                 ataactttactaaa
aataaaaccgcattaaaaaaactttattcaaATTTACGATCCAGAAATTTCCATAACGAATAACCTACTCATATAACGCC
TCttatattacatattattttaGGCTTATGCCGCCTGGAAAGATTGTTTAAGGCGCATTATATCGTGTCCGCTTTAACTA
TGGTacttaaactatttaaaatgtgGATTAATTTTACTGCACTATAAATATAGCACCCAAGACTACGAGATCTTAAAAAT
GTACGAGGACAAAATGAAACCTACAGGCCTAGATAACTAATGGCAAAAACCCCACAAACAAAAGATCTATTTAAACACCC
CTTACATATTTGGAACTTCCGTCATACATTGACTCGTGGTGGCAAGCTTGCTTATGGGCTAAGCGGCAATTGGGACCGCT
TGACCCTGGGAAGGTCTGAGACAATGACGTTAAATCTTCCGTGATGGGAGACGTAATAAATACGTTGGAAGGTACCAGAA
AACAACGGGATAATCGCTTGCGTAAGACCGAAAGCGACGCTATGCGGTTGTCCCCGGTTGAAAATCTTCCAGTCAGCGAT
TACAGTTCAATTAGCAAGGCGGCCCTTCGGTCGTAACTCAATTCAGACTACTCTGCAACCACAATTTTATtaatgctttg
tttttttgaccaCTAATTTAAGACCAGTCTAATTTTTCAATCATTTTTAGATCAATCtaaaaactctttttttagACCACT
CTGTGAAATTTTTCAgacacagtttaaaaaaaaaactttttttgttaaacatgtGTAAAATGACATTCAAATTTTCGGGCC
AGTGAAGAATCGTCCCTCTATGTTATCTGCTACCTACTAAAAGTAACCACCAACCACTGACCCCTAACTACCAGCAActt
ctaacactttccaccagcctaatctgtaacgtaccggaggattcttcactggtgcccttgaatgaatgaatgaatgaatg
aatgaatgaatgaatgaatgaatgaatgaatgaatgaatgaatgtaacgaaTGTGCCCTTGTAAGGTAATAAACTAACTG
TGGGCGAAATCGGAGTATTTTCTAGTCTGATCTTCTTATTATGACTCTAAAAGATCAAATATTTCTGaaccaacaaataa
aaatcaataaataatatatatttctacaaTACAGCATCAATTAAAACTTGGCAGATTGCCTTGGCACTTTATGGTCACAG
TGAACTACATGTGCAAGTACTAAACATAATAGTGGTAAGGGGTGGTAGCATTTCTATTACTTAAGTTTTCTGTAAAGTAA
TGCTTACACaagtaatgaataaatgaatgtagttTGCTTTATGTTCACGTAGCCAGAAAaaagacagtcattataacacg
ggtgttctgtttcatacacttcgtgcccgcttacaagttgctacatatataacttatttatcctcgcatggcgggggaAC
GAAAGTCTTTATAGAAaaggttctgtttcatacaccttgtgcccgcttataggttaccatgtatgttactctCTGAGTGA
TTTTTGGGGGAGTTTTTATGTAATACTTACATTACTTAACAGAAACACTTAGGTAGGTTGTAGATCTGTAACTTAACATT
AAGTAGAAAACATGCCGATCTTTAATTATTGTTGCCTTTCCTTAAATATATTCAAGTGGGCGCCTATAATTATAAGCGAC
GTACTTTAATCTCCcccttttaaacaaaaaaataagcattACTTAGCCTTAAGTAACAGAATCGAACTAACCAATGTGTT
AAGTATTACTAATGTAAGCAATACATAACGAAAAACTTAAGCAACAGAATTACATTCCATTCTGAGTGTCAATACAGTGA
ATCTAAACATAATTTCAAAGTCAGCACTAGCTGTAAAAATCTAAAGAACTATAGAATAATACTTAATCGATCAAAACTAG
AGGTTTAACAGTCTCTTCCGTGCccctttttttttgttatttttgctttttttgaaacattttcttgTTCCTTGCCTTGGG
AAATTTCCTGTGATTTATCCCCAGCTGCATTTGGCTCTTTAATTGTACTAATCAAAACCTGggagaaatttaaaatatag
tttgatATCAGCAAtggaatattaaacttttgctttttaagtagcaaaaaaaaaatttacaatttttttcgaAGTGTCAA
ACGTATAATGccagataaaaaataaaaaaaatatattgtgtttaatgTGGTCATACACCTTGTTTACATATCACTTACTA
ACTATGTAACTTGCTAACACAAGAATTAAGCTGGCATACCACTTCTTCAAAACACTCAGGAATAAAAGTGAGTGTAACGG
TCACACCTTTTACCcgaaacattatttctttttctgttttattaaaagcatgtttttttacaactaCTGTTCTACCagtc
```

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build KH
#!genome-build-accession NCBI_Assembly:GCF_000224145.2
##sequence-region NC_020166.2 1 10041005
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=7719
NC_020166.2 RefSeq  region  1    10041005    .   +   .    ID=id0;Dbxref=taxon
NC_020166.2 Gnomon  gene    2407    20500    .   -   .    ID=gene0;Dbxref=Gen
NC_020166.2 Gnomon  mRNA    2407    20500    .   -   .    ID=rna0;Parent=gene
NC_020166.2 Gnomon  exon    20332   20500    .   -   .    ID=id1;Parent=rna0;
NC_020166.2 Gnomon  exon    19895   19923    .   -   .    ID=id2;Parent=rna0;
NC_020166.2 Gnomon  exon    19077   19139    .   -   .    ID=id3;Parent=rna0;
NC_020166.2 Gnomon  exon    18475   18606    .   -   .    ID=id4;Parent=rna0;
NC_020166.2 Gnomon  exon    17884   17988    .   -   .    ID=id5;Parent=rna0;
NC_020166.2 Gnomo                                          d6;Parent=rna0;
NC_020166.2 Gnomo                                          d7;Parent=rna0;
NC_020166.2 Gnomo                                          d8;Parent=rna0;
NC_020166.2 Gnomo                                          d9;Parent=rna0;
NC_020166.2 Gnomon  exon    15049   15175    .   -   .    ID=id10;Parent=rna0
NC_020166.2 Gnomon  exon    15011   15046    .   -   .    ID=id11;Parent=rna0
NC_020166.2 Gnomon  exon    14512   14637    .   -   .    ID=id12;Parent=rna0
NC_020166.2 Gnomon  exon    13778   13971    .   -   .    ID=id13;Parent=rna0
NC_020166.2 Gnomon  exon    13416   13483    .   -   .    ID=id14;Parent=rna0
NC_020166.2 Gnomon  exon    12989   13061    .   -   .    ID=id15;Parent=rna0
NC_020166.2 Gnomon  exon    12572   12674    .   -   .    ID=id16;Parent=rna0
NC_020166.2 Gnomon  exon    12094   12214    .   -   .    ID=id17;Parent=rna0
NC_020166.2 Gnomon  exon    11669   11830    .   -   .    ID=id18;Parent=rna0
NC_020166.2 Gnomon  exon    11040   11223    .   -   .    ID=id19;Parent=rna0
NC_020166.2 Gnomon  exon    10640   10847    .   -   .    ID=id20;Parent=rna0
NC_020166.2 Gnomon  exon    10242   10308    .   -   .    ID=id21;Parent=rna0
NC_020166.2 Gnomon  exon    9992    10100    .   -   .    ID=id22;Parent=rna0
NC_020166.2 Gnomon  exon    9394    9459     .   -   .    ID=id23;Parent=rna0
NC_020166.2 Gnomon  exon    9081    9182     .   -   .    ID=id24;Parent=rna0
NC_020166.2 Gnomon  exon    8556    8750     .   -   .    ID=id25;Parent=rna0
NC_020166.2 Gnomon  exon    8106    8186     .   -   .    ID=id26;Parent=rna0
NC_020166.2 Gnomon  exon    7558    7721     .   -   .    ID=id27;Parent=rna0
NC_020166.2 Gnomon  exon    7033    7195     .   -   .    ID=id28;Parent=rna0
NC_020166.2 Gnomon  exon    6244    6471     .   -   .    ID=id29;Parent=rna0
NC_020166.2 Gnomon  exon    5728    5889     .   -   .    ID=id30;Parent=rna0
NC_020166.2 Gnomon  exon    5300    5479     .   -   .    ID=id31;Parent=rna0
NC_020166.2 Gnomon  exon    4841    5005     .   -   .    ID=id32;Parent=rna0
NC_020166.2 Gnomon  exon    4179    4407     .   -   .    ID=id33;Parent=rna0
NC_020166.2 Gnomon  exon    3781    3884     .   -   .    ID=id34;Parent=rna0
NC_020166.2 Gnomon  exon    2407    3338     .   -   .    ID=id35;Parent=rna0
NC_020166.2 Gnomon  CDS 20332   20356    .   -   0   ID=cds0;Parent=rna0;Dbx
```

Topics

- Comparative genomics: orthologs/paralogs prediction
- **Comparative genomics: functional annotation**
- Structural proteomics: 3D-modeling of protein structures

Functional annotation

UniProtKB

UniProt Knowledgebase

Swiss-Prot (557,713)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (116,030,110)

Automatically annotated and not reviewed.

Records that await full manual annotation.

Gene Ontology Consortium

KEGG Kyoto Encyclopedia of Genes and Genomes

InterPro

Protein sequence analysis & classification

Pfam

ProDom

PANTHER Classification System

prosite

Conserved Domains

reactome

HAMAP

Structure Function Linkage Database

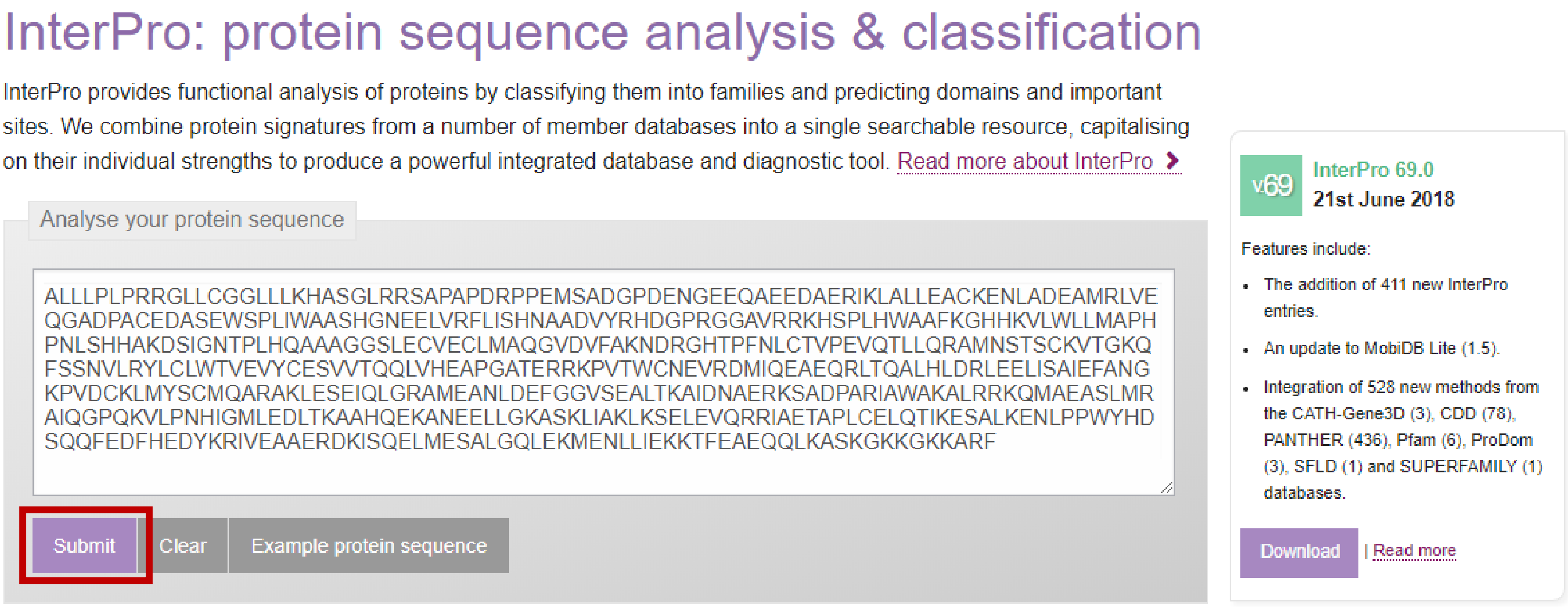
Superfamily 1.75

HMM library and genome assignments server

SMART

CATH / Gene3D v4.2

Functional annotation



InterPro: protein sequence analysis & classification

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. Read more about InterPro

Analyse your protein sequence

```
ALLLPLPRRGLLCGGLLLKHASGLRRSAPAPDRPPEMSADGPDENGEEQAEEDAERIKLALLEACKENLADEAMRLVE
QGADPACEDASEWSPLIWAASHGNEELVRFLISHNAADVYRHDGPRGGAVRRKHSPLHWAAFKGHHKVLWLLMAPH
PNLSHHAKDSIGNTPLHQAAAGGSLECVECLMAQGVDVFAKNDRGHTPFNLCTVPEVQTLLQRAMNSTSCKVTGKQ
FSSNVLRYLCLWTVEVYCESVVTQQLVHEAPGATERRKPVTWCNEVRDMIQEAEQRLTQALHLDRLEELISAIEFANG
KPVDCKLMYSCMQARAKLESEIQLGRAMEANLDEFGGVSEALTKAIDNAERKSADPARIAWAKALRRKQMAEASLMR
AIQGPQKVLPNHIGMLEDLTKAAHQEKANEELLGKASKLIAKLKSELEVQRRIAETAPLCELQTIKESALKENLPPWYHD
SQQFEDFHEDYKRIVEAAERDKISQELMESALGQLEKMENLLIEKKTFEAEQQLKASKGKKGKKARF
```

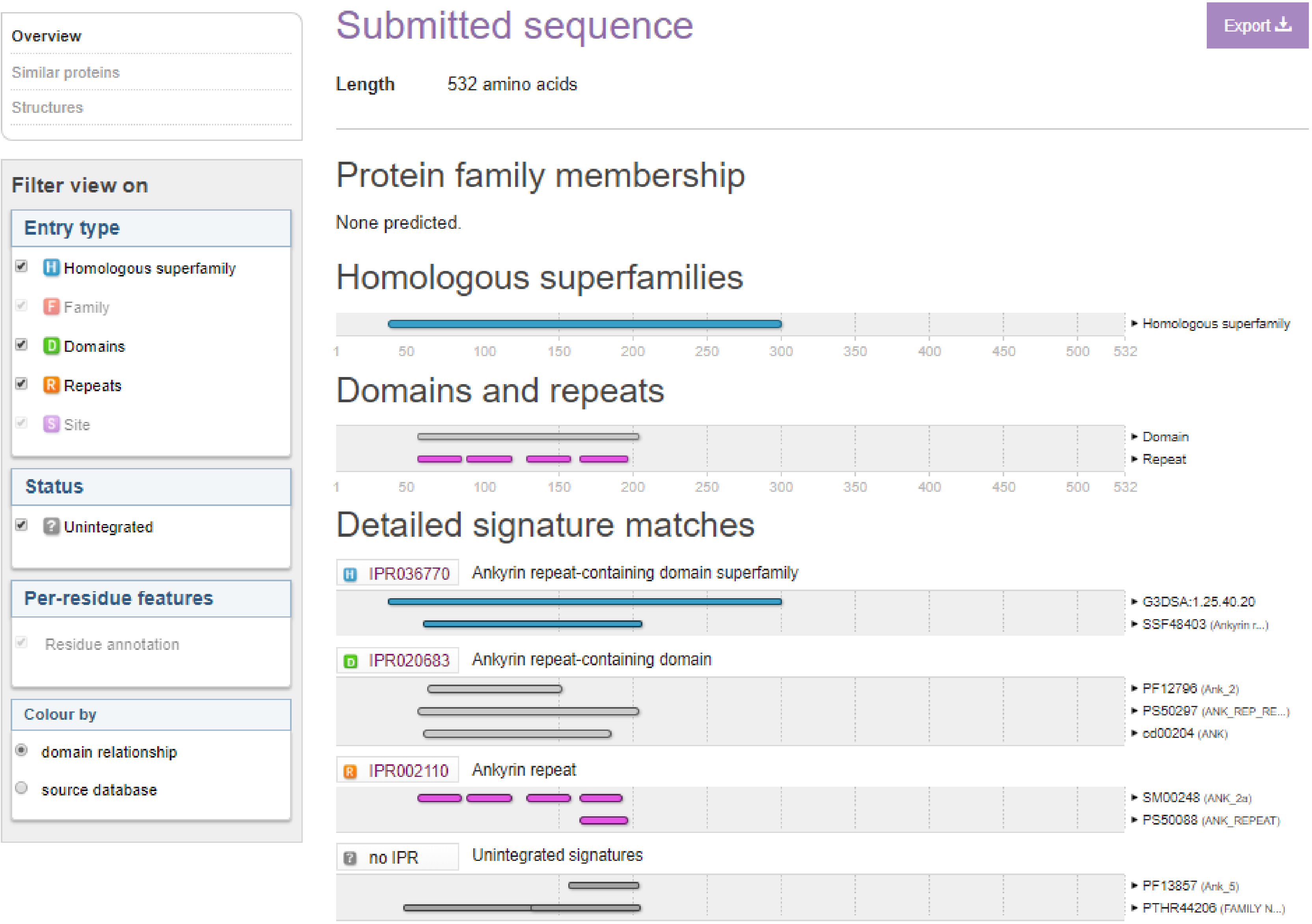
v69 InterPro 69.0

21st June 2018

Features include:

- The addition of 411 new InterPro entries.
- An update to MobiDB Lite (1.5).
- Integration of 528 new methods from the CATH-Gene3D (3), CDD (78), PANTHER (436), Pfam (6), ProDom (3), SFLD (1) and SUPERFAMILY (1) databases.

Functional annotation



Overview

Similar proteins

Structures

Filter view on

Entry type

☑ H Homologous superfamily

☑ F Family

☑ D Domains

☑ R Repeats

☑ S Site

Status

☑ ? Unintegrated

Per-residue features

☑ Residue annotation

Colour by

● domain relationship

○ source database

Submitted sequence

Length 532 amino acids

Protein family membership

None predicted.

Homologous superfamilies

Domains and repeats

Detailed signature matches

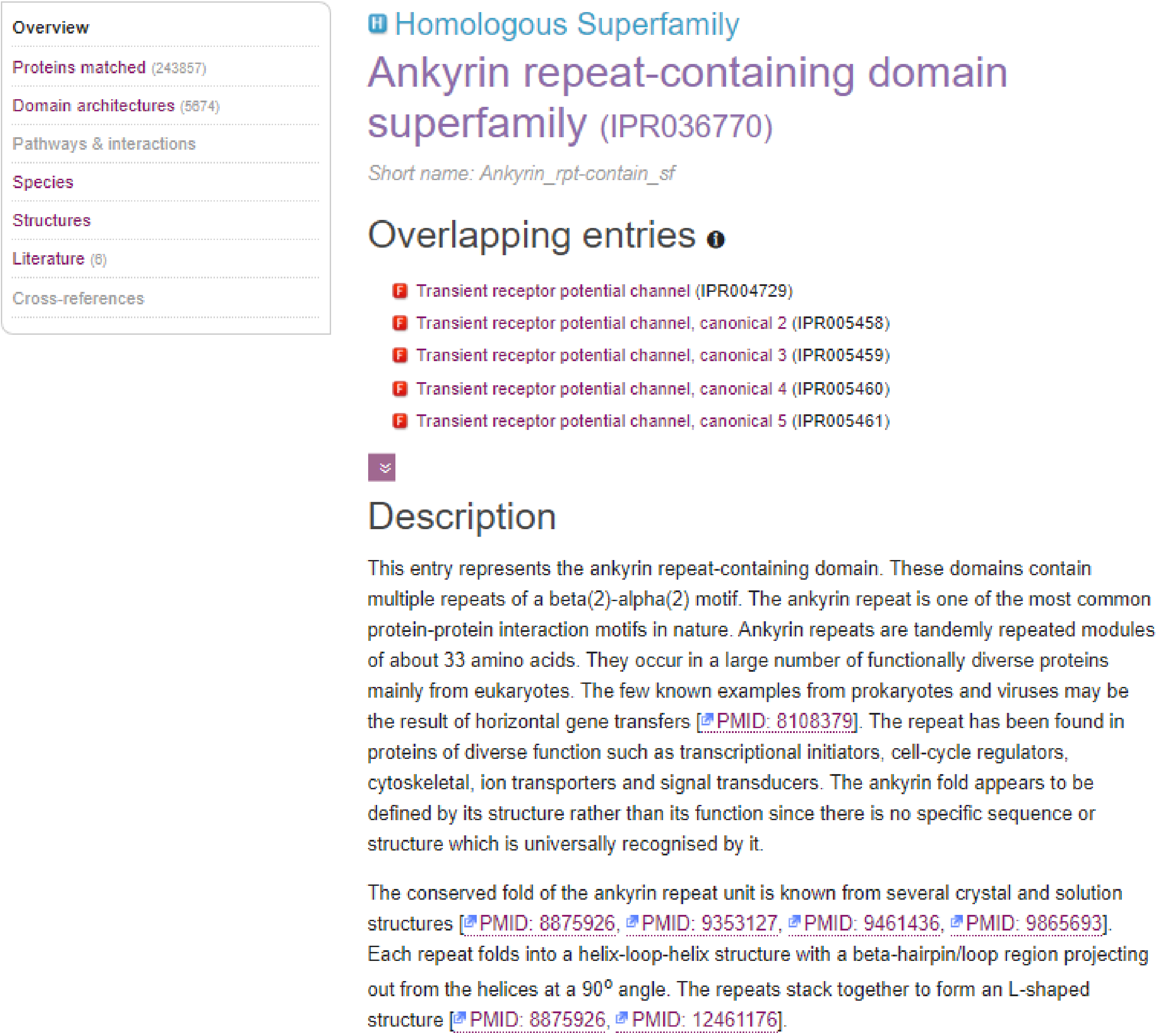
H IPR036770 Ankyrin repeat-containing domain superfamily

D IPR020683 Ankyrin repeat-containing domain

R IPR002110 Ankyrin repeat

? no IPR Unintegrated signatures

Functional annotation



- Overview
- Proteins matched (243857)
- Domain architectures (5674)
- Pathways & interactions
- Species
- Structures
- Literature (6)
- Cross-references

H Homologous Superfamily

Ankyrin repeat-containing domain superfamily (IPR036770)

Short name: Ankyrin_rpt-contain_sf

Overlapping entries

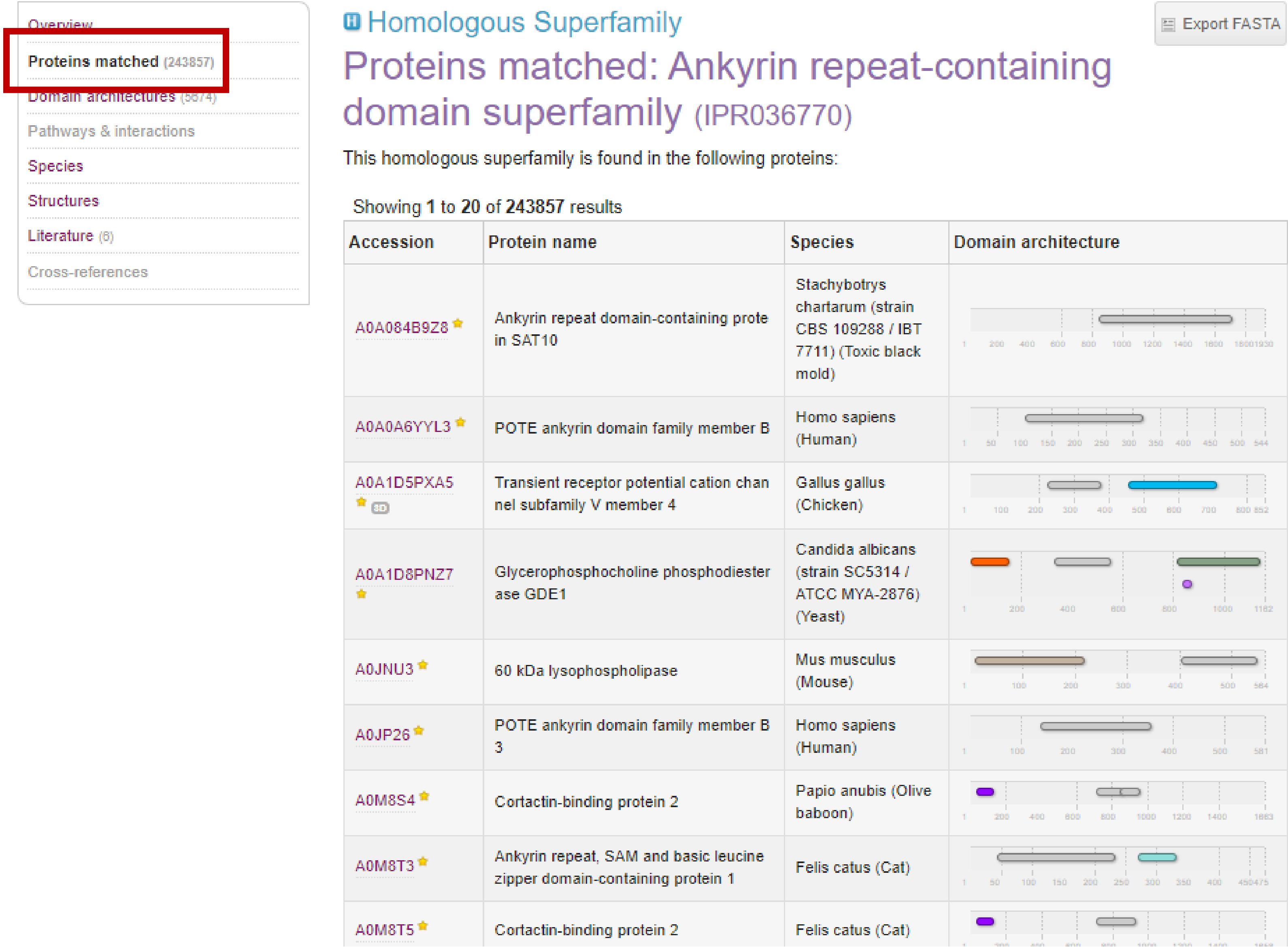
- F Transient receptor potential channel (IPR004729)
- F Transient receptor potential channel, canonical 2 (IPR005458)
- F Transient receptor potential channel, canonical 3 (IPR005459)
- F Transient receptor potential channel, canonical 4 (IPR005460)
- F Transient receptor potential channel, canonical 5 (IPR005461)

Description

This entry represents the ankyrin repeat-containing domain. These domains contain multiple repeats of a beta(2)-alpha(2) motif. The ankyrin repeat is one of the most common protein-protein interaction motifs in nature. Ankyrin repeats are tandemly repeated modules of about 33 amino acids. They occur in a large number of functionally diverse proteins mainly from eukaryotes. The few known examples from prokaryotes and viruses may be the result of horizontal gene transfers [PMID: 8108379]. The repeat has been found in proteins of diverse function such as transcriptional initiators, cell-cycle regulators, cytoskeletal, ion transporters and signal transducers. The ankyrin fold appears to be defined by its structure rather than its function since there is no specific sequence or structure which is universally recognised by it.

The conserved fold of the ankyrin repeat unit is known from several crystal and solution structures [PMID: 8875926, PMID: 9353127, PMID: 9461436, PMID: 9865693]. Each repeat folds into a helix-loop-helix structure with a beta-hairpin/loop region projecting out from the helices at a 90° angle. The repeats stack together to form an L-shaped structure [PMID: 8875926, PMID: 12461176].

Functional annotation



- Overview
- **Proceins matched** (243857)
- Domain architectures (5674)
- Pathways & interactions
- Species
- Structures
- Literature (6)
- Cross-references

Export FASTA

Homologous Superfamily

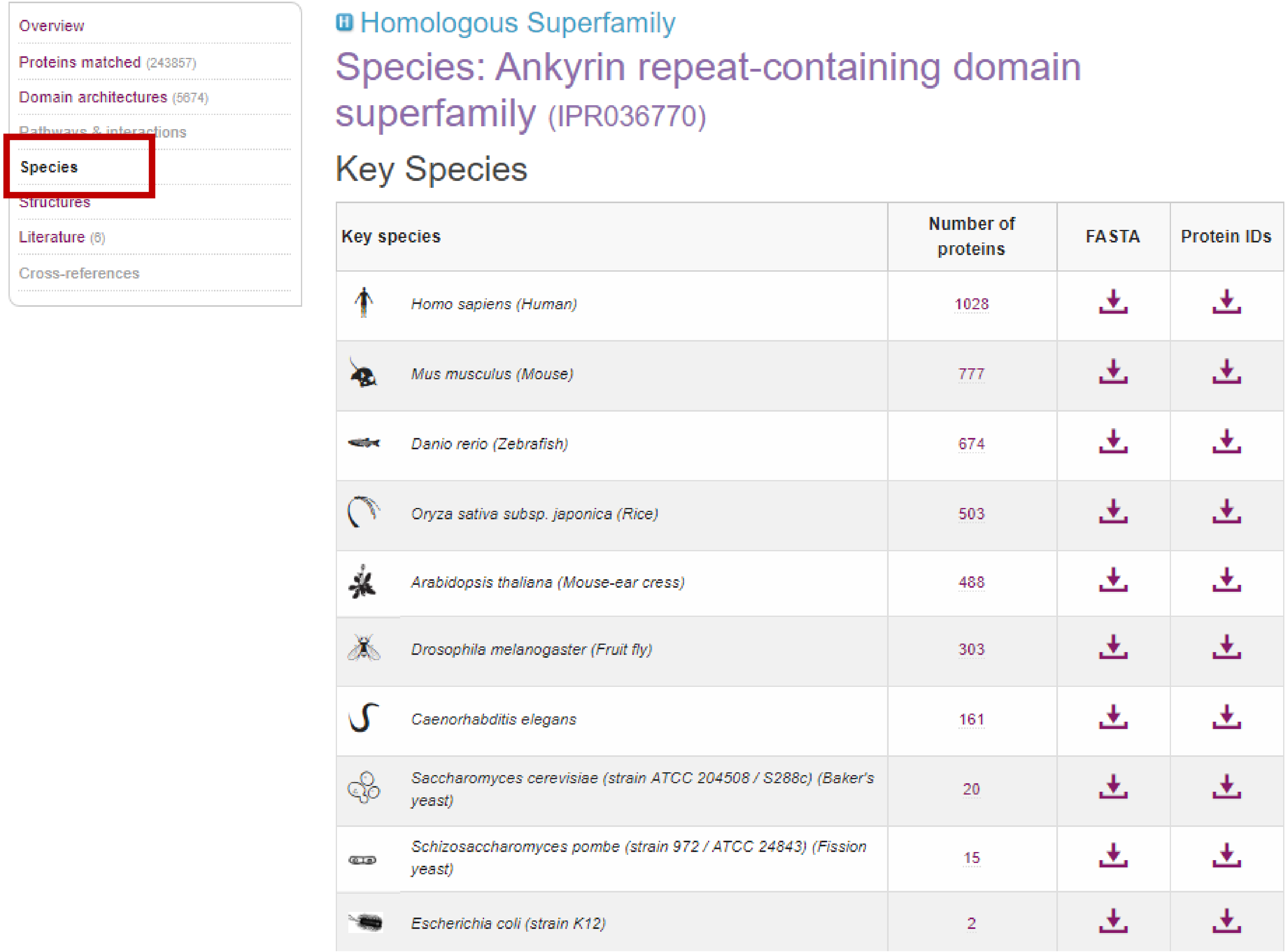
Proteins matched: Ankyrin repeat-containing domain superfamily (IPR036770)

This homologous superfamily is found in the following proteins:

Showing **1** to **20** of **243857** results

Accession	Protein name	Species	Domain architecture
A0A084B9Z8	Ankyrin repeat domain-containing protein SAT10	Stachybotrys chartarum (strain CBS 109288 / IBT 7711) (Toxic black mold)	
A0A0A6YYL3	POTE ankyrin domain family member B	Homo sapiens (Human)	
A0A1D5PXA5	Transient receptor potential cation channel subfamily V member 4	Gallus gallus (Chicken)	
A0A1D8PNZ7	Glycerophosphocholine phosphodiesterase GDE1	Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast)	
A0JNU3	60 kDa lysophospholipase	Mus musculus (Mouse)	
A0JP26	POTE ankyrin domain family member B 3	Homo sapiens (Human)	
A0M8S4	Cortactin-binding protein 2	Papio anubis (Olive baboon)	
A0M8T3	Ankyrin repeat, SAM and basic leucine zipper domain-containing protein 1	Felis catus (Cat)	
A0M8T5	Cortactin-binding protein 2	Felis catus (Cat)	

Functional annotation



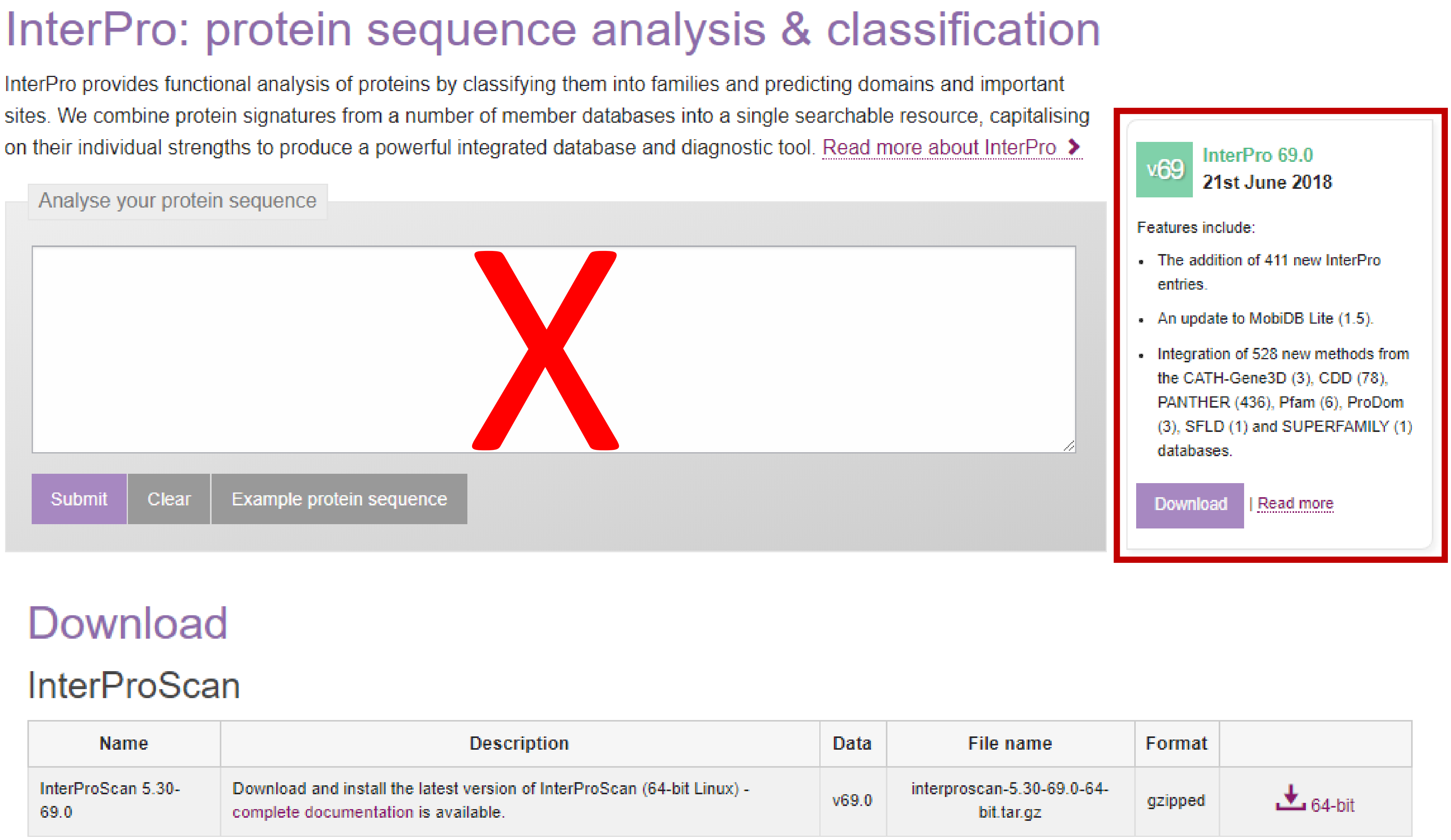
Homologous Superfamily

Species: Ankyrin repeat-containing domain superfamily (IPR036770)

Key Species

Key species	Number of proteins	FASTA	Protein IDs
Homo sapiens (Human)	1028	⬇	⬇
Mus musculus (Mouse)	777	⬇	⬇
Danio rerio (Zebrafish)	674	⬇	⬇
Oryza sativa subsp. japonica (Rice)	503	⬇	⬇
Arabidopsis thaliana (Mouse-ear cress)	488	⬇	⬇
Drosophila melanogaster (Fruit fly)	303	⬇	⬇
Caenorhabditis elegans	161	⬇	⬇
Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	20	⬇	⬇
Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)	15	⬇	⬇
Escherichia coli (strain K12)	2	⬇	⬇

Functional annotation



InterPro: protein sequence analysis & classification

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. Read more about InterPro

v69 InterPro 69.0
21st June 2018

Features include:

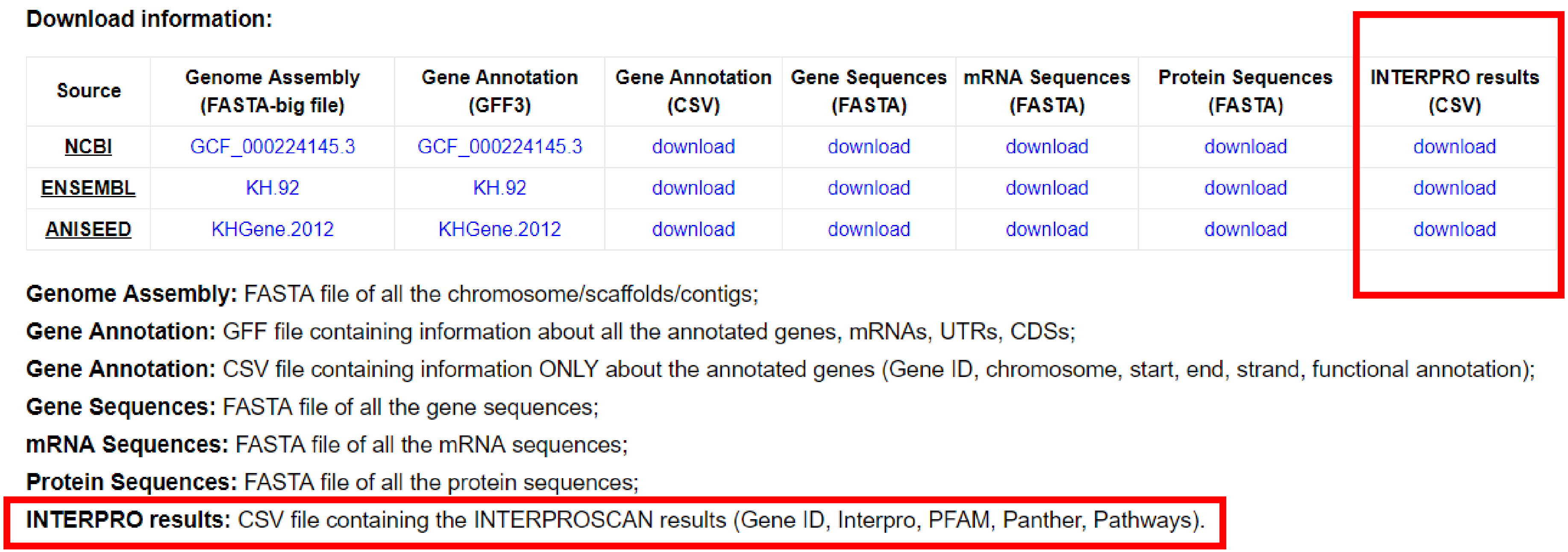
- The addition of 411 new InterPro entries.
- An update to MobiDB Lite (1.5).
- Integration of 528 new methods from the CATH-Gene3D (3), CDD (78), PANTHER (436), Pfam (6), ProDom (3), SFLD (1) and SUPERFAMILY (1) databases.

Download

InterProScan

Name	Description	Data	File name	Format	
InterProScan 5.30-69.0	Download and install the latest version of InterProScan (64-bit Linux) - complete documentation is available.	v69.0	interproscan-5.30-69.0-64-bit.tar.gz	gzipped	64-bit

The GENOMA platform



Download information:

Source	Genome Assembly (FASTA-big file)	Gene Annotation (GFF3)	Gene Annotation (CSV)	Gene Sequences (FASTA)	mRNA Sequences (FASTA)	Protein Sequences (FASTA)	INTERPRO results (CSV)
NCBI	GCF_000224145.3	GCF_000224145.3	download	download	download	download	download
ENSEMBL	KH.92	KH.92	download	download	download	download	download
ANISEED	KHGene.2012	KHGene.2012	download	download	download	download	download

Genome Assembly: FASTA file of all the chromosome/scaffolds/contigs;

Gene Annotation: GFF file containing information about all the annotated genes, mRNAs, UTRs, CDSs;

Gene Annotation: CSV file containing information ONLY about the annotated genes (Gene ID, chromosome, start, end, strand, functional annotation);

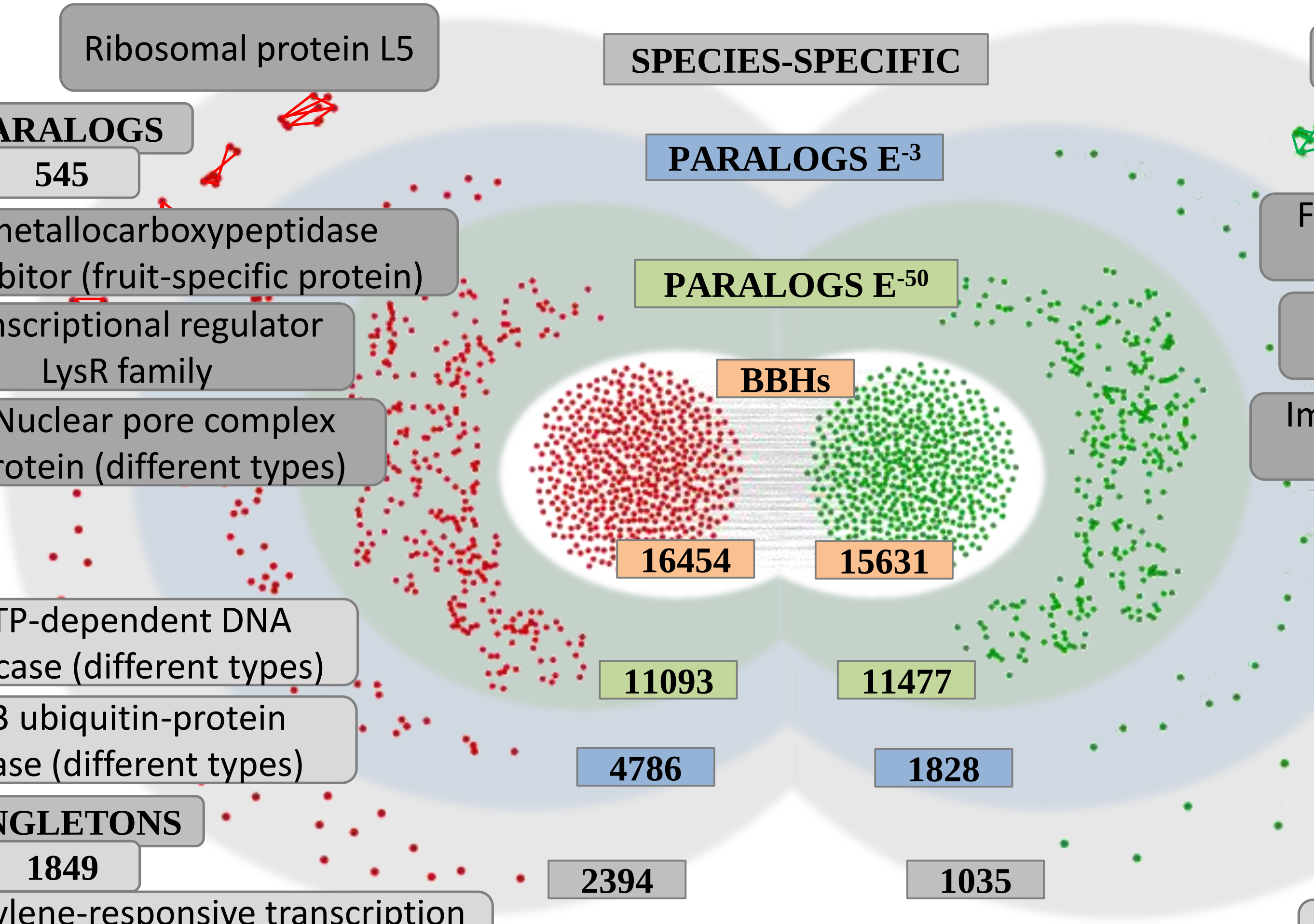
Gene Sequences: FASTA file of all the gene sequences;

mRNA Sequences: FASTA file of all the mRNA sequences;

Protein Sequences: FASTA file of all the protein sequences;

INTERPRO results: CSV file containing the INTERPROSCAN results (Gene ID, Interpro, PFAM, Panther, Pathways).

Functional annotation



SPECIES-SPECIFIC

PARALOGS E^{-3}

PARALOGS E^{-50}

BBHs

16454	15631
11093	11477
4786	1828
2394	1035

PARALOGS
545

Ribosomal protein L5

metallocarboxypeptidase inhibitor (fruit-specific protein)

Transcriptional regulator LysR family

Nuclear pore complex protein (different types)

ATP-dependent DNA helicase (different types)

E3 ubiquitin-protein ligase (different types)

SINGLETONS
1849

Ethylene-responsive transcription factor (different types)

Zinc finger domains (different types)

peptidylprolyl isomerase EC: 5.2.1.8

PARALOGS
107

Agglutinin

Fanconi anaemia protein FANCD2

GPI ethanolamine phosphate transferase 2

Immediate early response 3-interacting protein 1

DNA repair protein RAD50

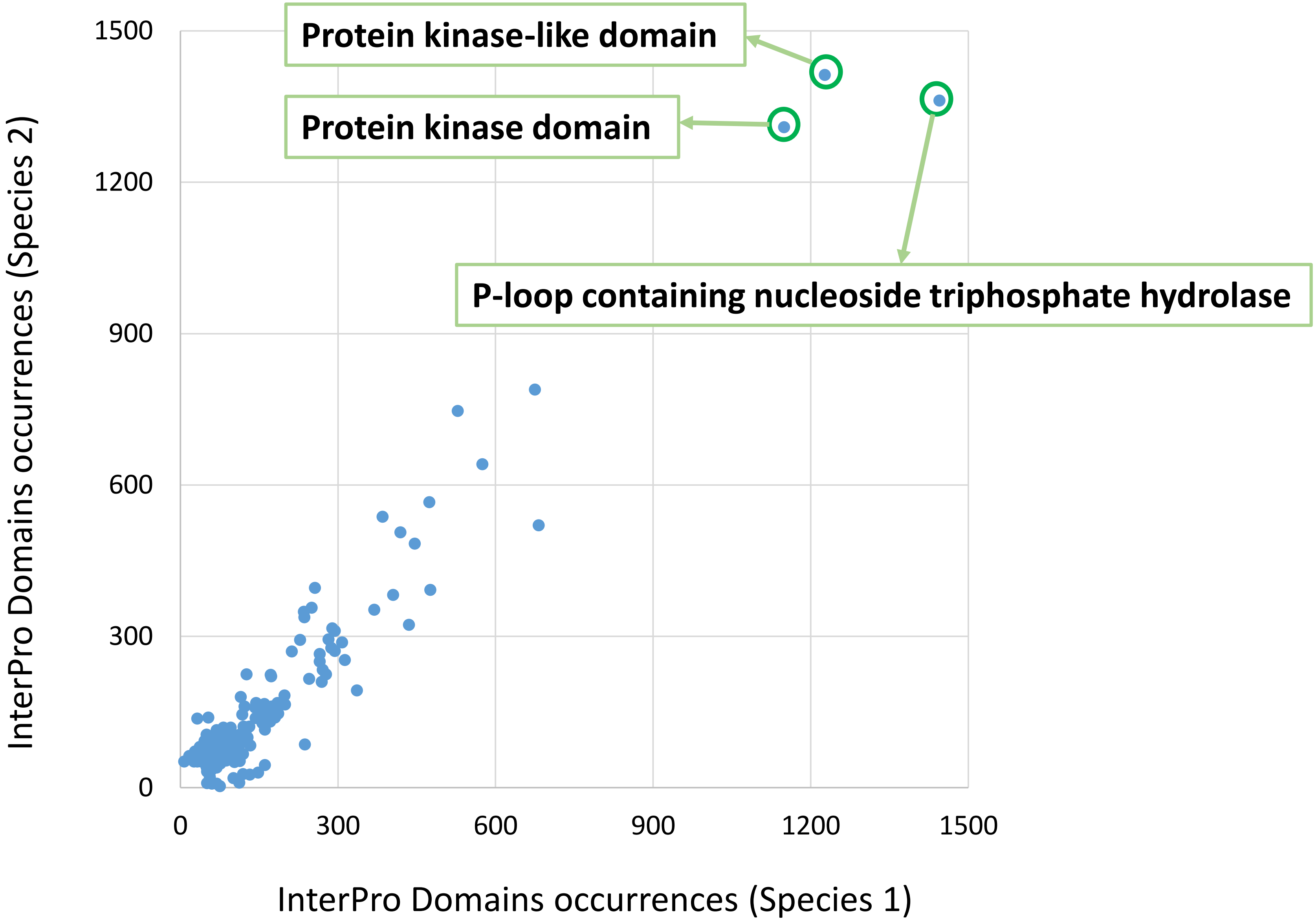
SINGLETONS
928

Metallothionein-like protein type 3

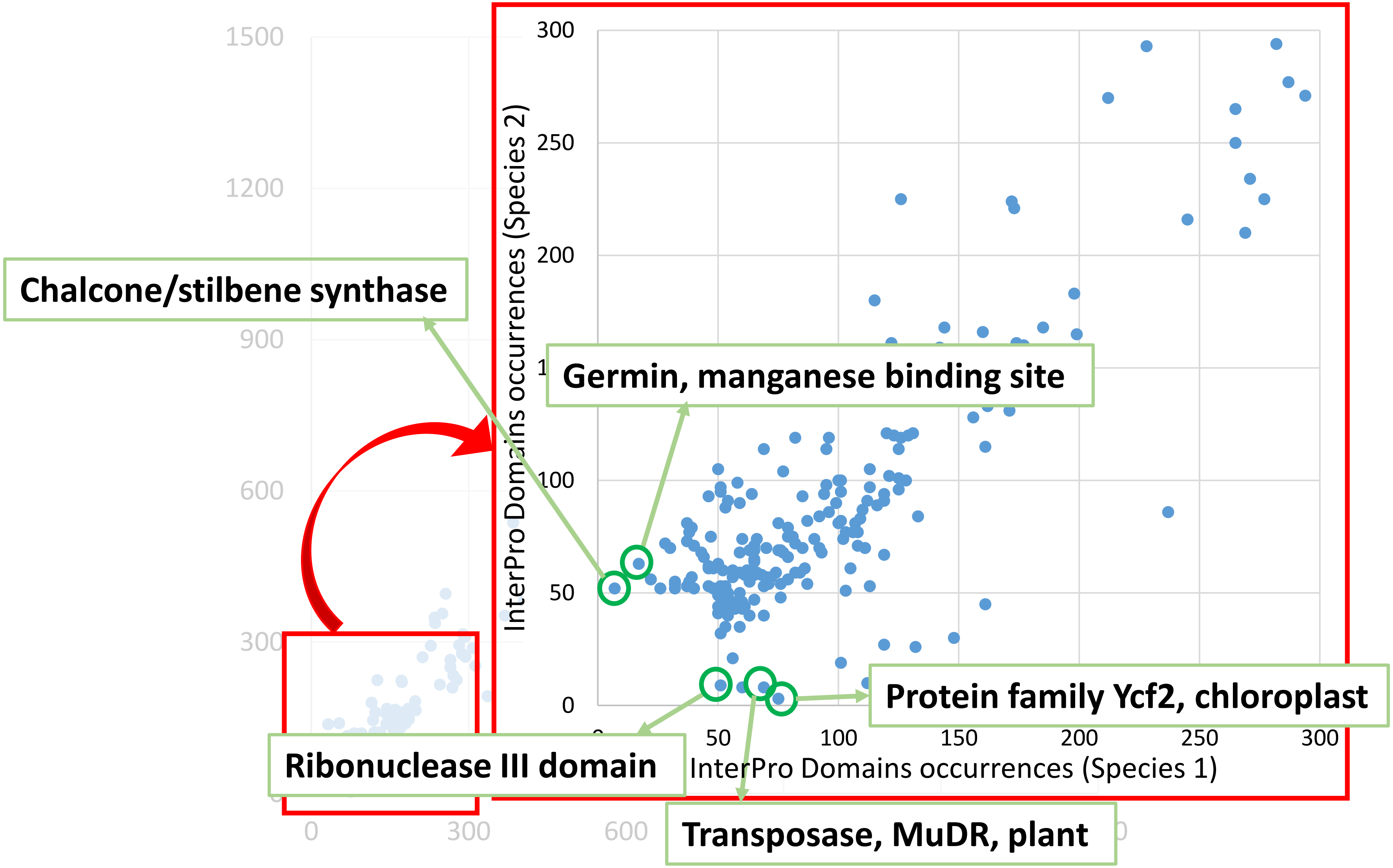
Rubredoxin-type fold domain

● Spe

Functional annotation



Functional annotation



Chalcone/stilbene synthase

Germin, manganese binding site

Protein family Ycf2, chloroplast

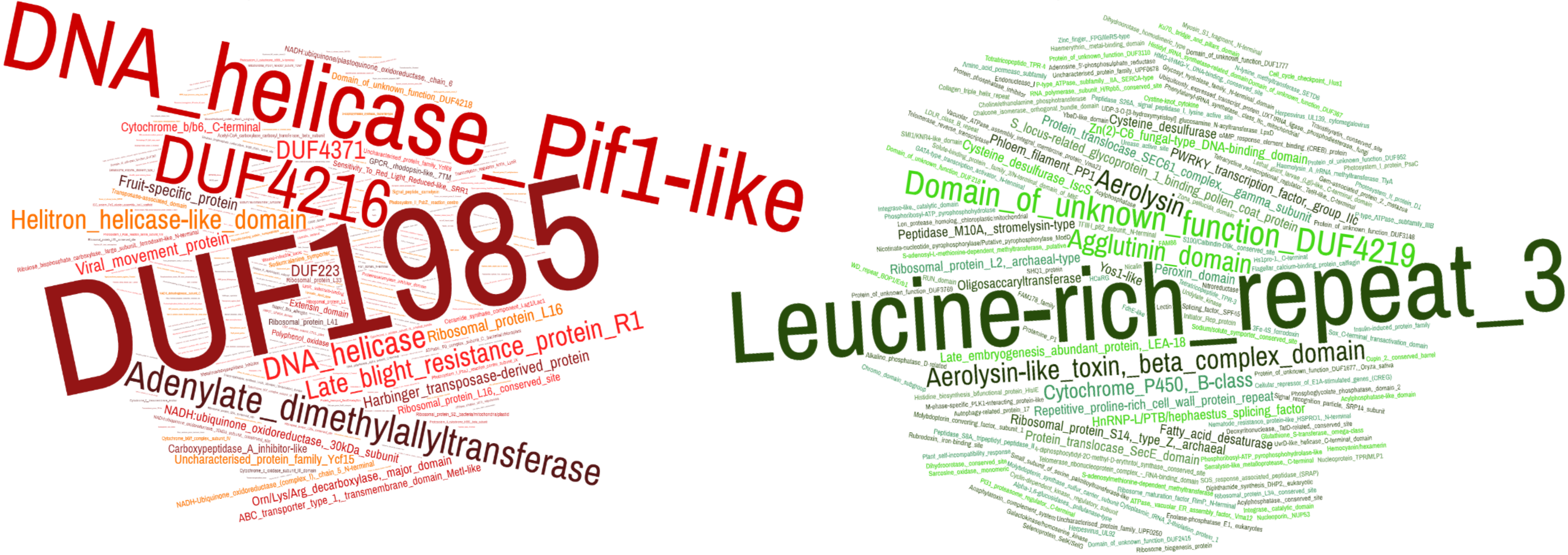
Ribonuclease III domain

Transposase, MuDR, plant

Functional annotation

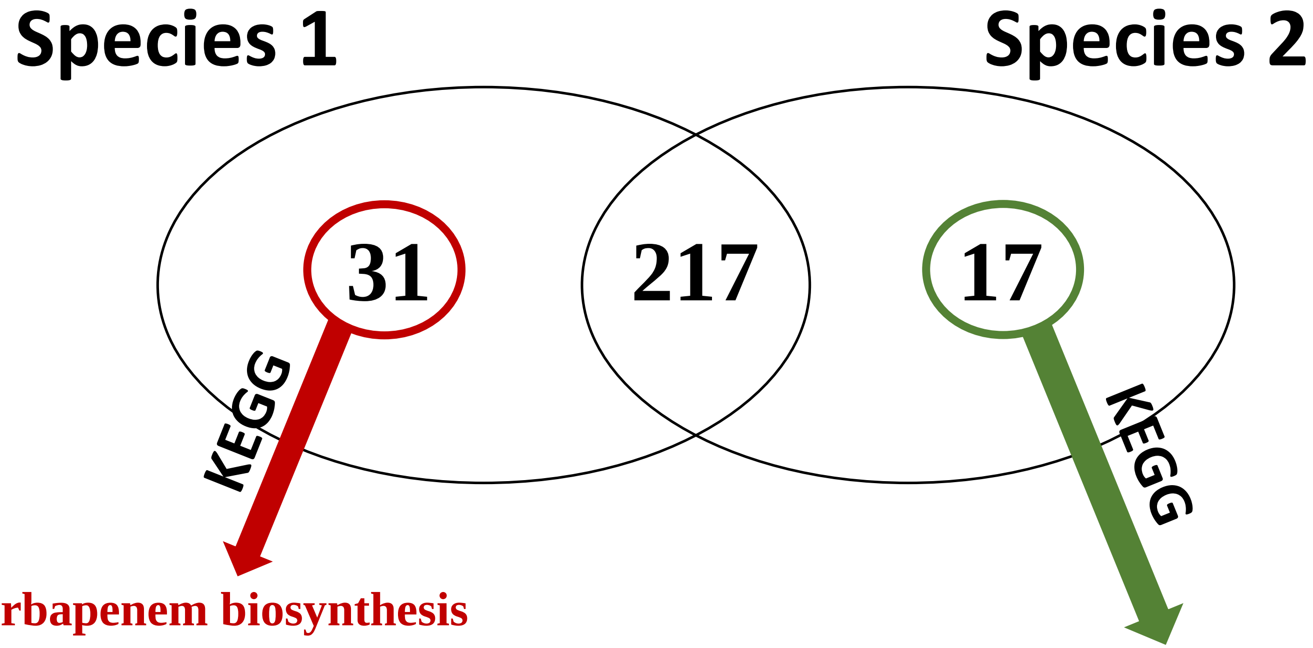
Species 1

Species 2



SPECIES-SPECIFIC DOMAINS

Functional annotation



Carbapenem biosynthesis

Ethylbenzene degradation

Aflatoxin biosynthesis

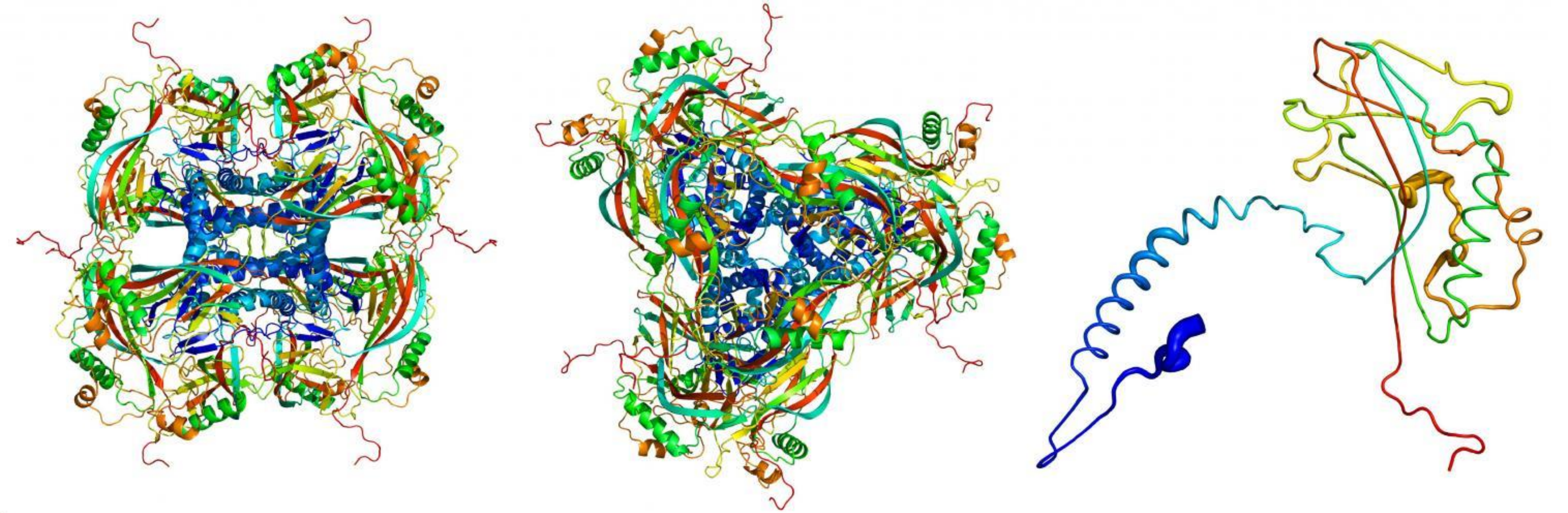
Monoterpenoid biosynthesis

Nicotinate and nicotinamide metabolism

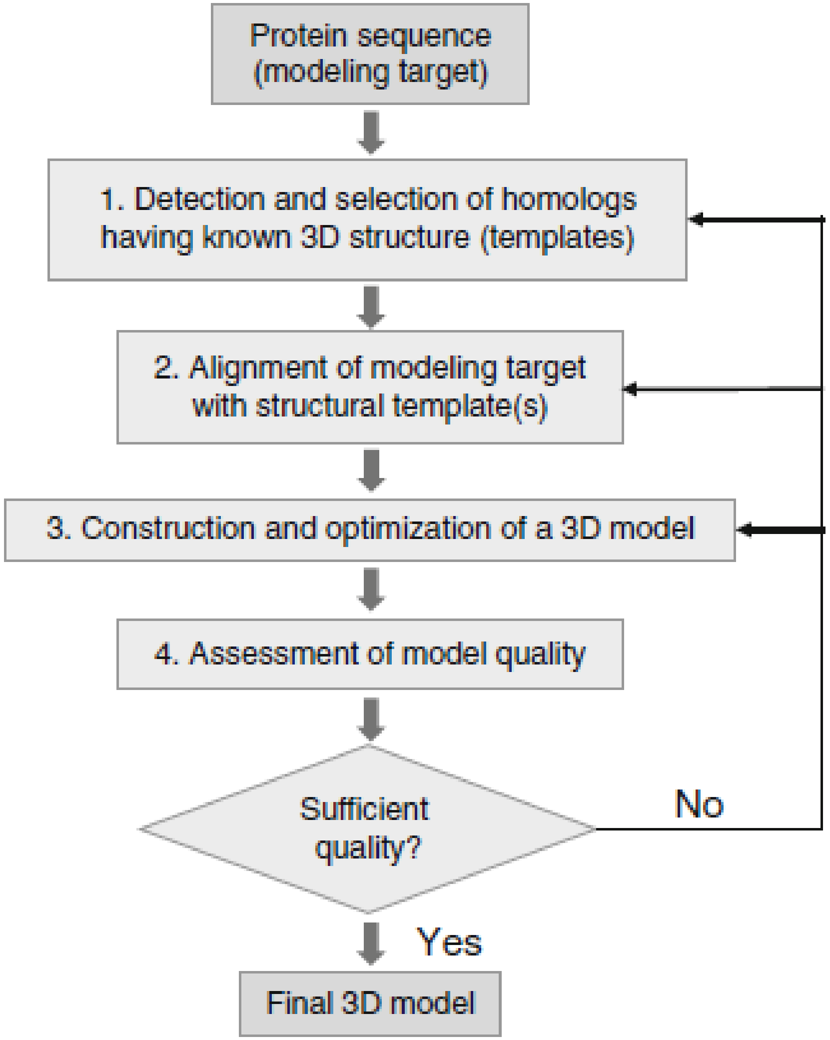
Glycosaminoglycan biosynthesis
(heparan sulfate / heparin)

Topics

- Comparative genomics: orthologs/paralogs prediction
- Comparative genomics: functional annotation
- **Structural proteomics: 3D-modeling of protein structures**



Homology modeling


Protein sequence (modeling target)

1. Detection and selection of homologs having known 3D structure (templates)

2. Alignment of modeling target with structural template(s)

3. Construction and optimization of a 3D model

4. Assessment of model quality

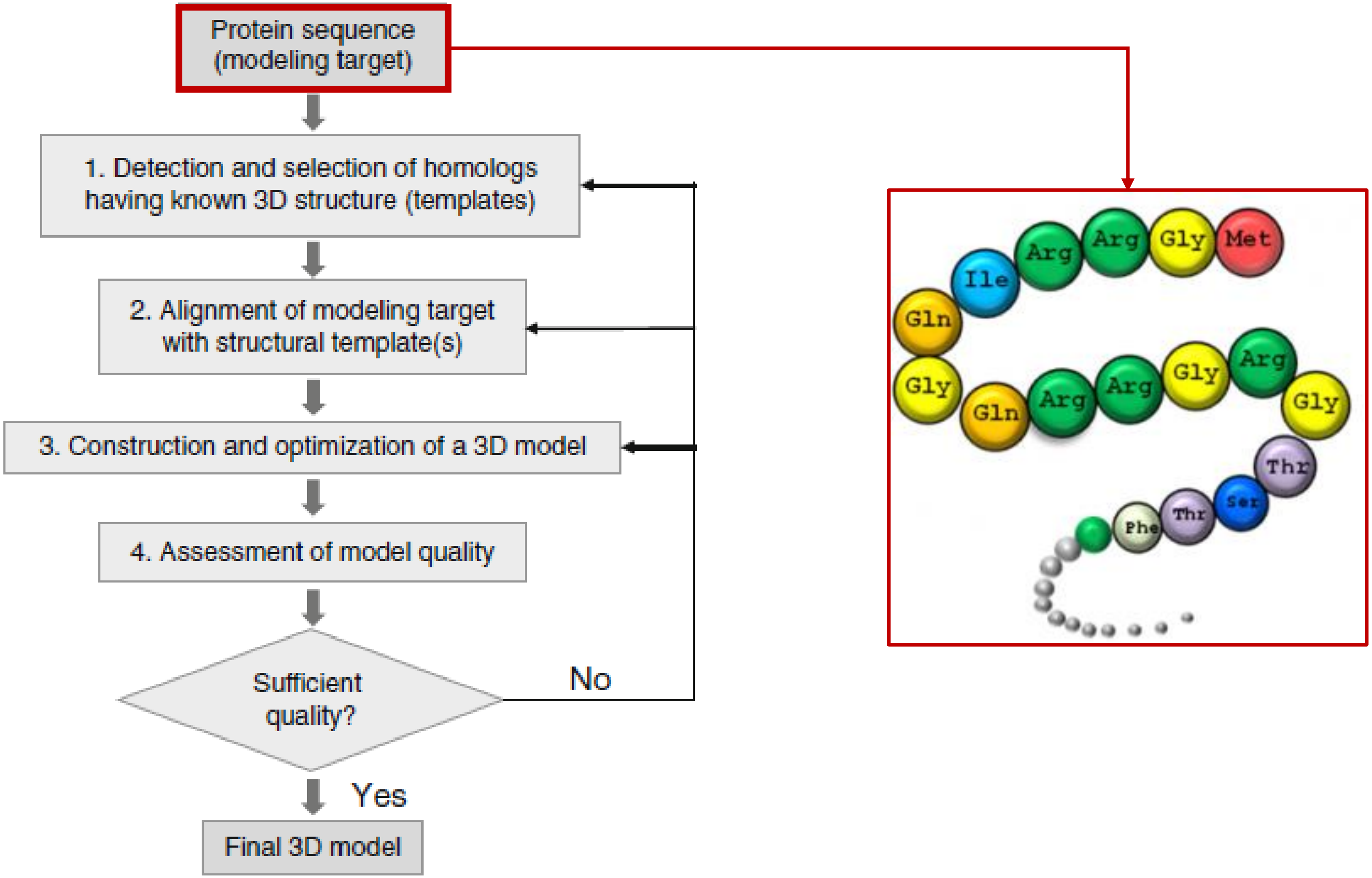
Sufficient quality?

No

Yes

Final 3D model

Homology modeling



Protein sequence (modeling target)

1. Detection and selection of homologs having known 3D structure (templates)

2. Alignment of modeling target with structural template(s)

3. Construction and optimization of a 3D model

4. Assessment of model quality

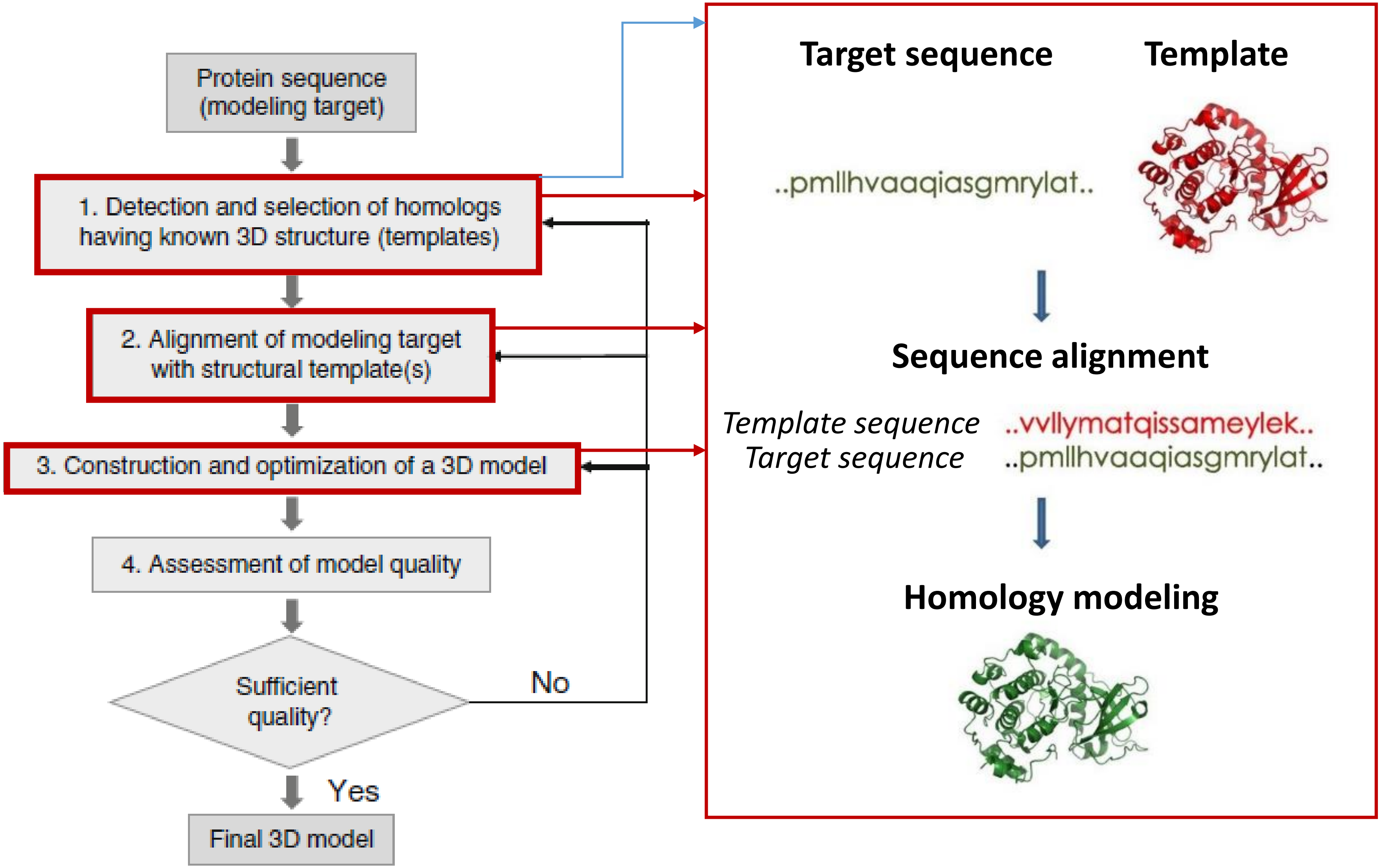
Sufficient quality?

No

Yes

Final 3D model

Homology modeling



Protein sequence (modeling target)

1. Detection and selection of homologs having known 3D structure (templates)
2. Alignment of modeling target with structural template(s)
3. Construction and optimization of a 3D model
4. Assessment of model quality

Sufficient quality?

No

Yes

Final 3D model

Target sequence **Template**

..pmllhvaaqiasgmrylat..

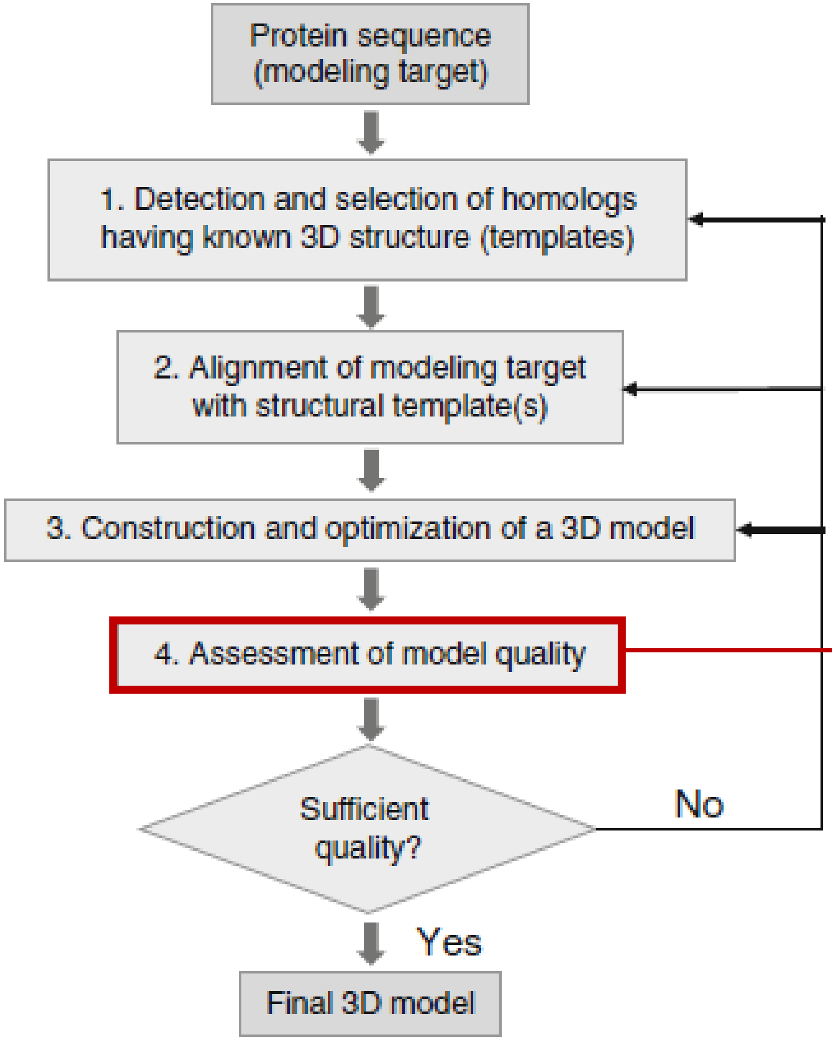
Sequence alignment

Template sequence ..vvllymatqissameylek..

Target sequence ..pmllhvaaqiasgmrylat..

Homology modeling

Homology modeling



Protein sequence (modeling target)

1. Detection and selection of homologs having known 3D structure (templates)

2. Alignment of modeling target with structural template(s)

3. Construction and optimization of a 3D model

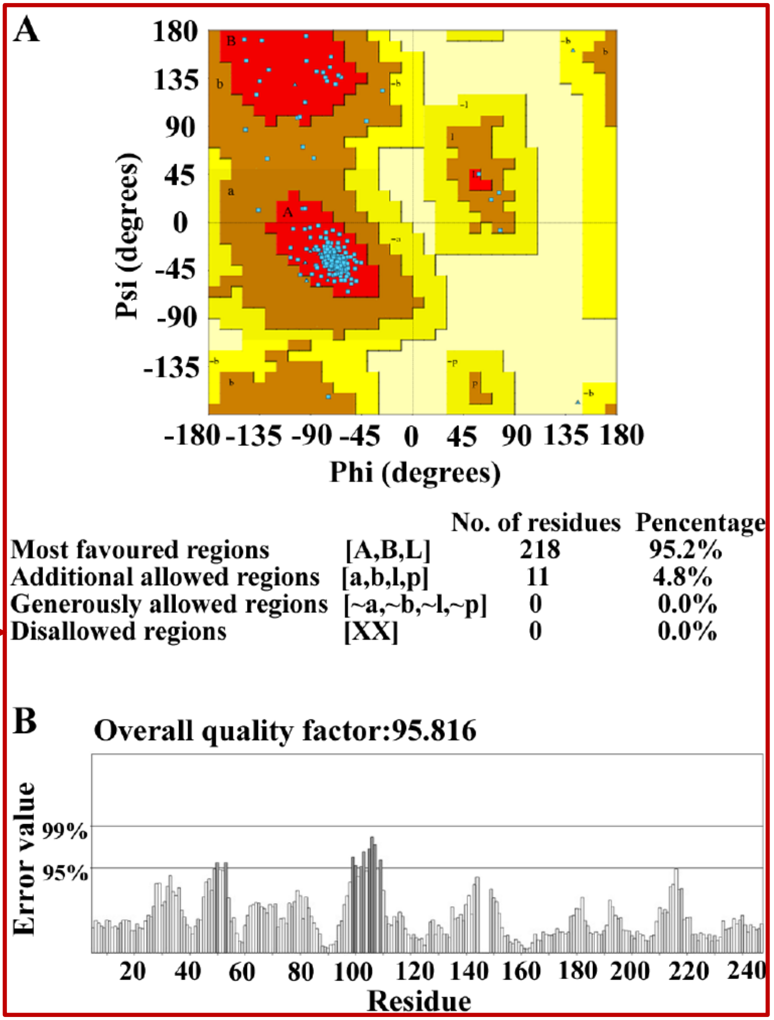
4. Assessment of model quality

Sufficient quality?

No

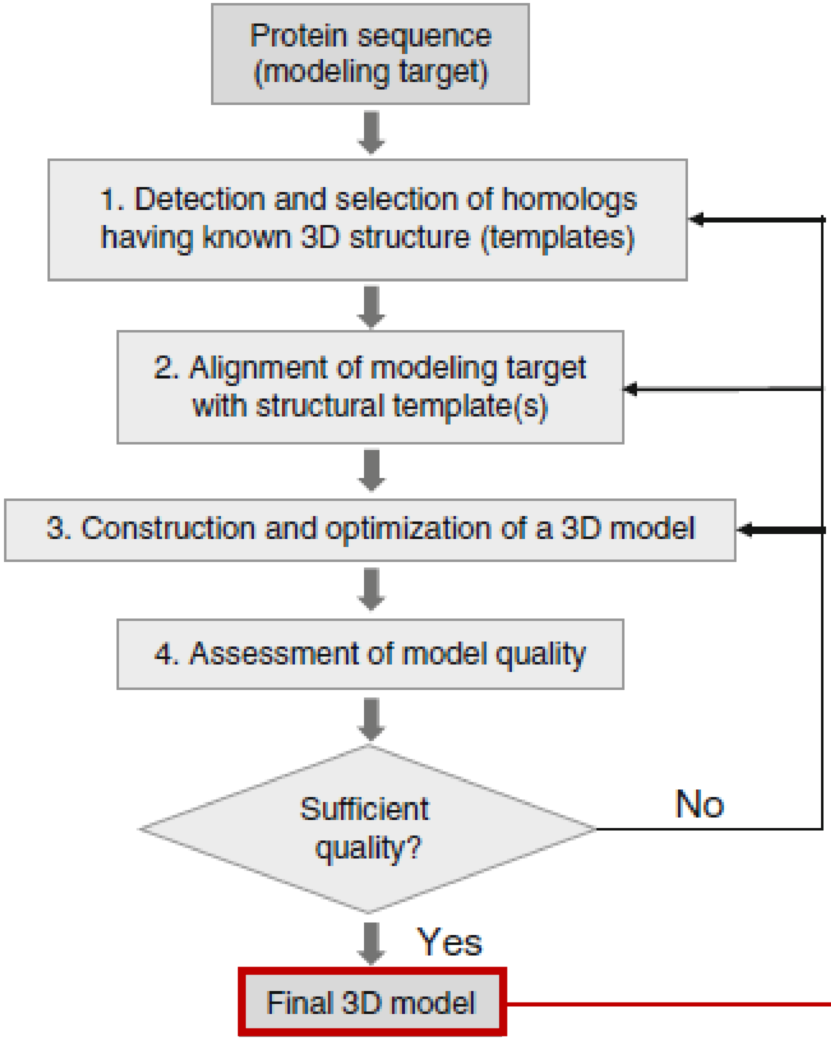
Yes

Final 3D model



		No. of residues	Pencentage
Most favoured regions	[A,B,L]	218	95.2%
Additional allowed regions	[a,b,l,p]	11	4.8%
Generously allowed regions	[~a,~b,~l,~p]	0	0.0%
Disallowed regions	[XX]	0	0.0%

Homology modeling



Protein sequence (modeling target)

1. Detection and selection of homologs having known 3D structure (templates)

2. Alignment of modeling target with structural template(s)

3. Construction and optimization of a 3D model

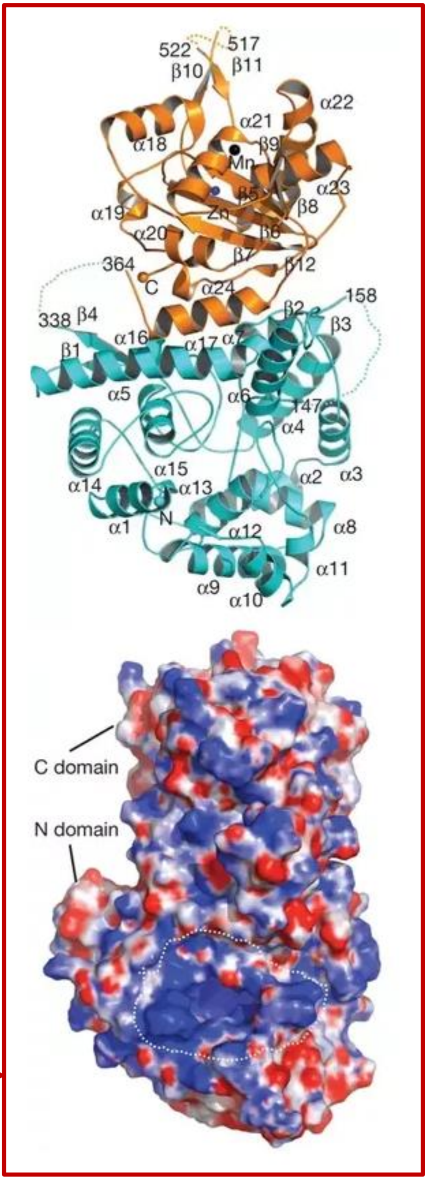
4. Assessment of model quality

Sufficient quality?

No

Yes

Final 3D model



Conclusions

Already available services

- ✓ Orthologs, paralogs and species-specific genes
- ✓ Functional annotations
- ✓ 3-D modeling

Soon available

- COMPARO: **COM**parative for **PAR**alog and **O**rtolog genes
- Molecular dynamics simulations