Annotation & Inference
New genomes, New functions

Having Function

Experiments
Literature
Expert view

'Maybe'

Boarder line similarity
Only part of protein
Conflicting exp/ lit

'Wrong'

Fault annotation
Wrong inference

No Function

New genomes
No similarity
No evidence

Michal Linial ,Institute of Life Sciences

The Hebrew University of Jerusalem

Annotation & Inference
New genomes, New functions

Domain families by EVEREST
Automatic identification of Protein Domain
Performance and analysis w.r.t to other resources

New Annotation by Inference
A method for inference – testing on a new genome

New Function to Disserted Proteins
High level functionality – story of the toxin like proteins

Why domain families?

what is wrong with protein classification

Nothing is wrong, But:

- Reducing **false** transitivity.
- Exposing **Mix and Match** evolution
- **Immediate relevance** to **structural** domain-families
- Suggesting evolutionary '**robust units**'

Why automatic?

Overcoming large **amounts of data**

Unbiased identification of new families (even without an identified seed)

EVEREST : A domain families resource
A comparative quality tool for other resources

Automatic / de-novo identification and classification of protein domains in all known sequences

Rigorous evaluation against manually / automated

& structurally based domain- family resources

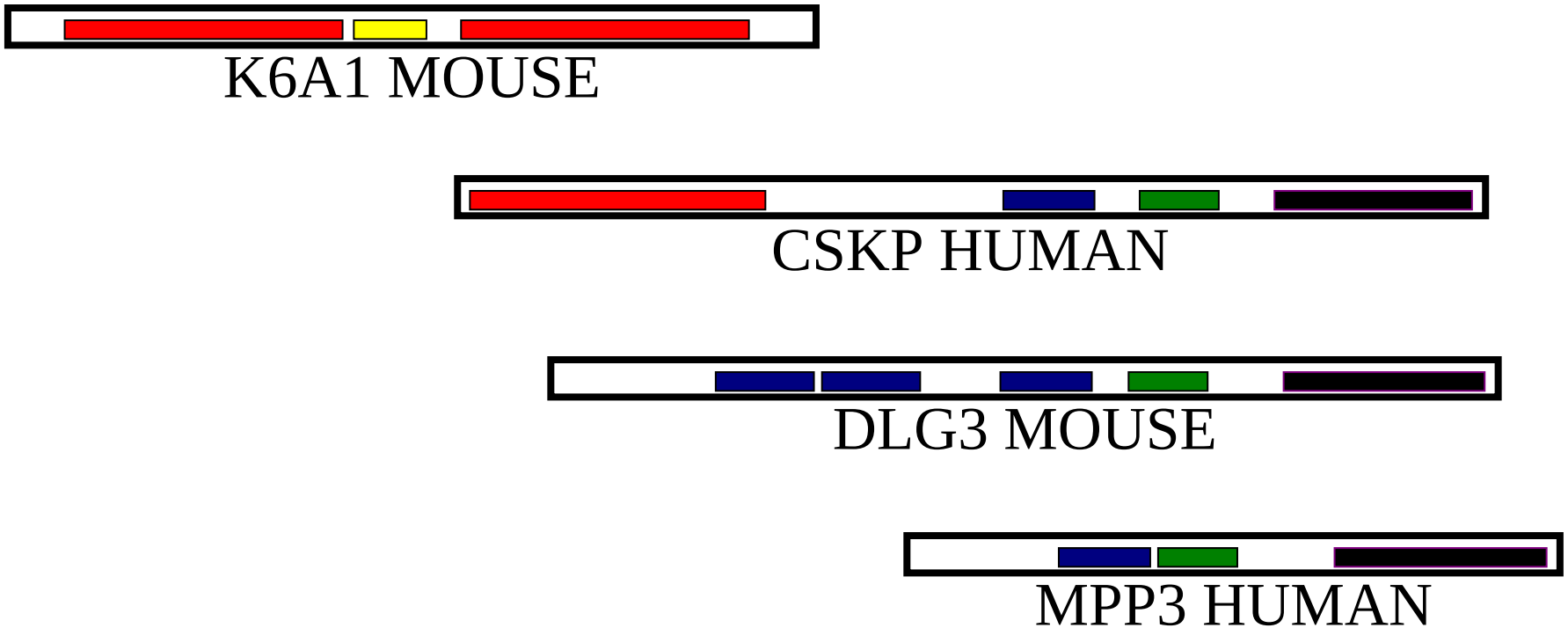
- Scoring methods for a **'quality control'**
- Exposing any (interesting) relationships within 'the world' of domains

- **Web interactive tool**

  www.everest.cs.huji.ac.il

Method

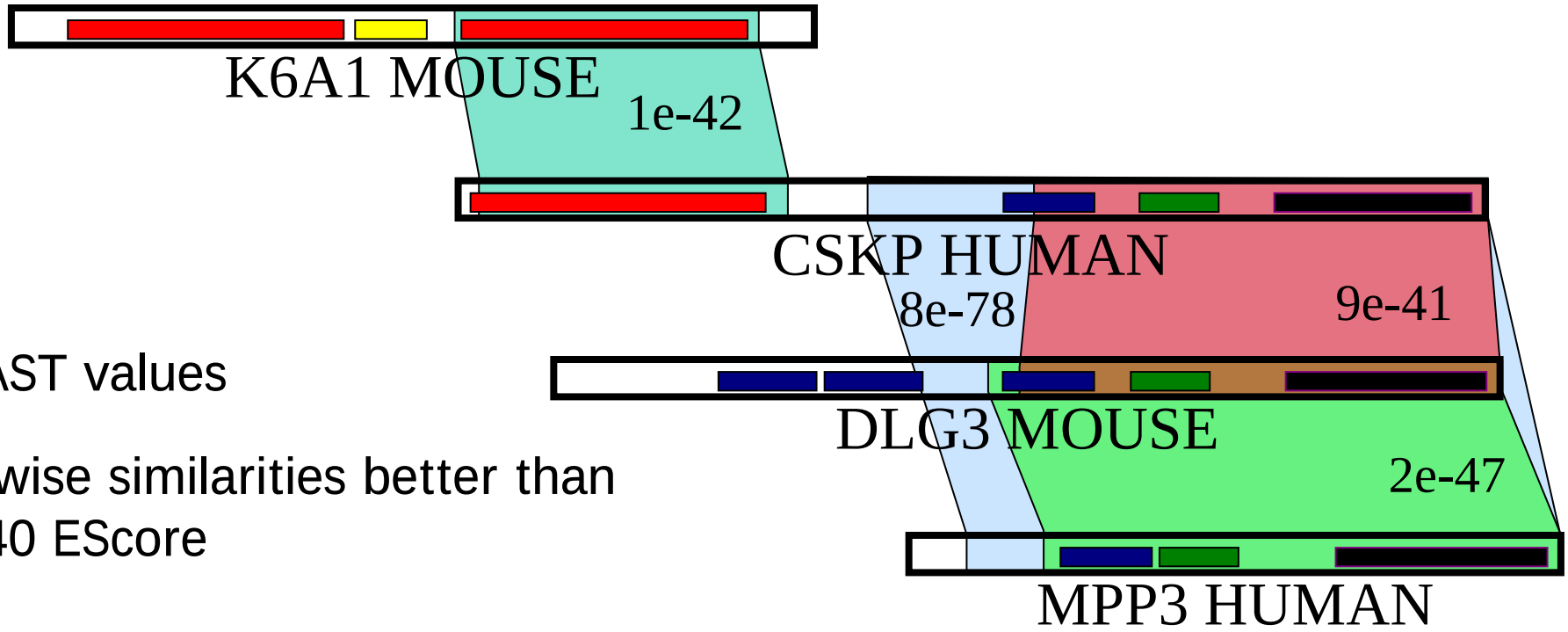
The Modular Nature of Proteins



- Serine/Threonine protein kinase family active site
- Protein kinase C-terminal domain
- PDZ domain
- SH3 domain
- Guanylate kinase

input

False Transitivity of Local Alignment



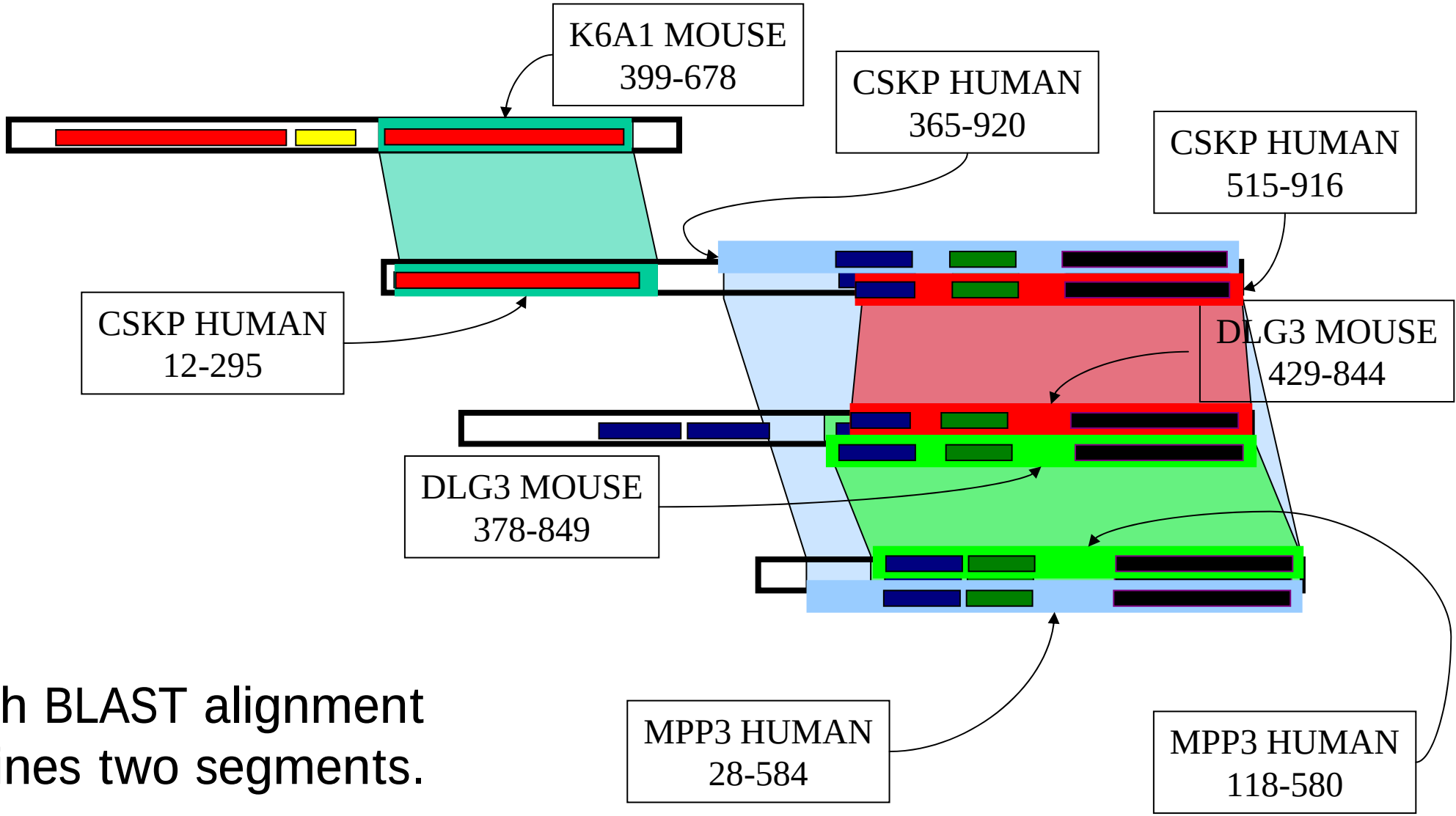
BLAST values

Pairwise similarities better than 1e-40 EScore

If we cluster these proteins, assuming transitivity of local alignment scores, we will cluster K6A1_MOUSE with MPP3_HUMAN

Method

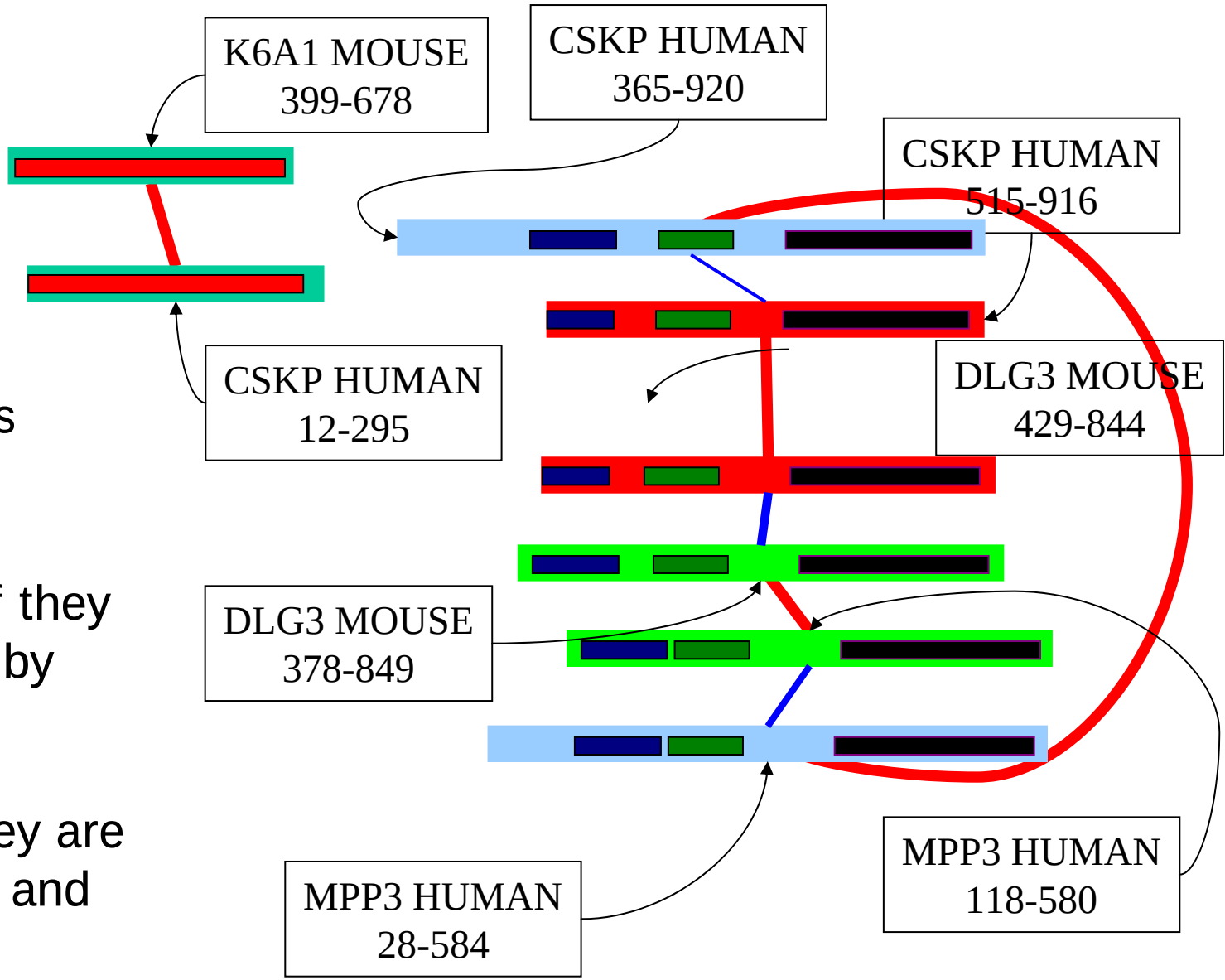
Working With Segments



Each BLAST alignment defines two segments.

input

Clustering Segments



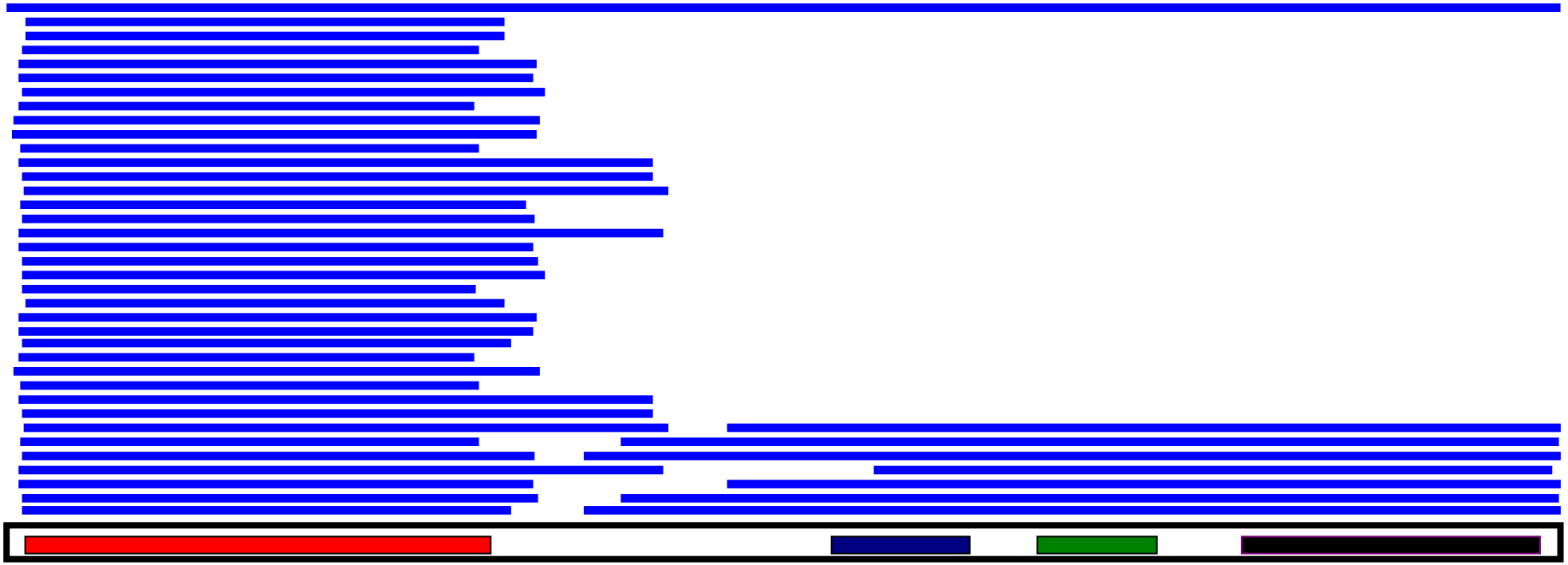
Two similarity measures between segments:

- Sequence similarity if they were found together by BLAST
- Physical overlap if they are on the same protein, and they intersect

input

The Easy Case

All segments on CSKP_HUMAN defined by alignments with e-score 1e-40 or better:



CSKP HUMAN

We collect all Blast value that are < 100 ! ~14 million values

EVEREST: Process Scheme

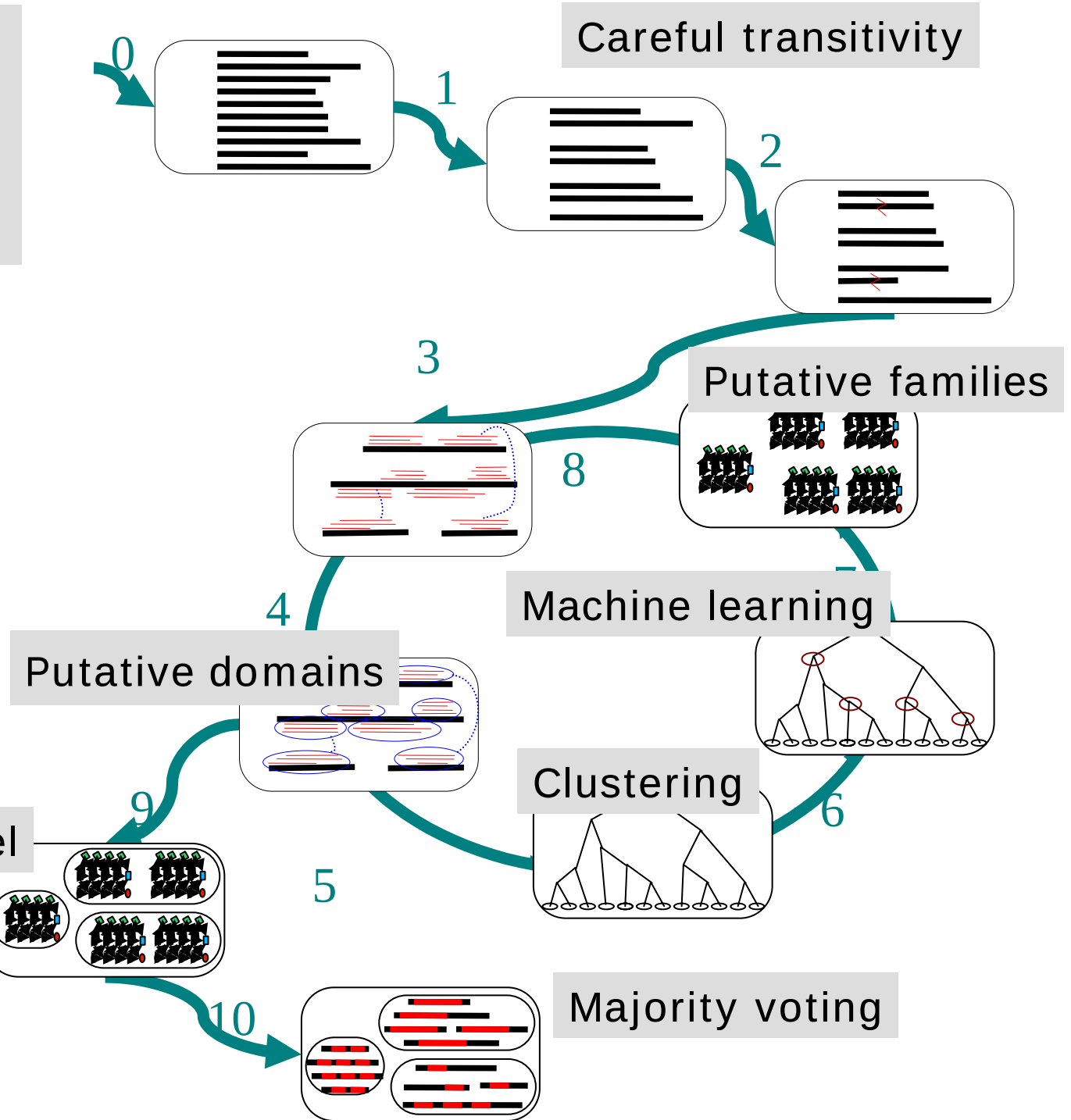
EVolutionary
Ensembles of
REcurrent SegmenTs

Pre-process

Iterations

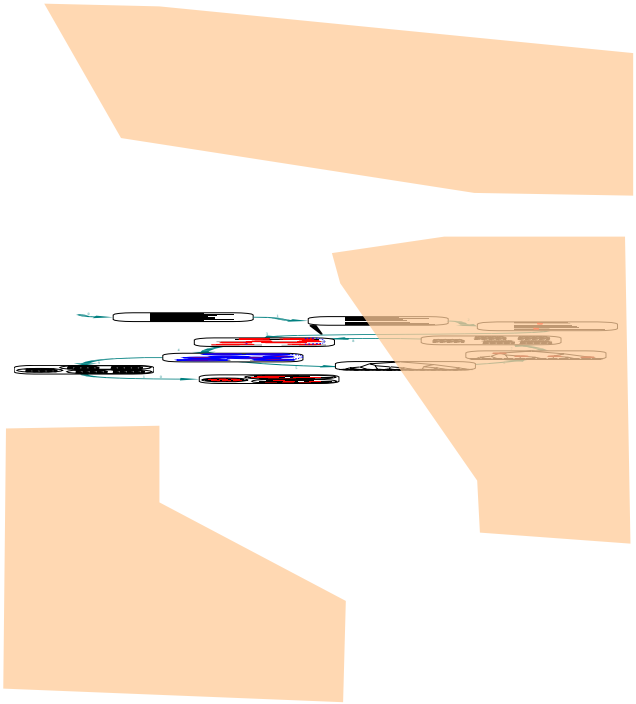
post-process

Evaluation and tests

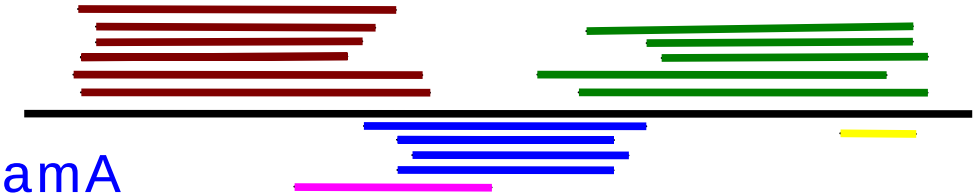


Method

3 Years in one slide
(Elon Portugaly)



- Cluster the segments into conservative groups by overlap similarity
- Each group is a **putative domain**
- We apply average linkage hierarchical clustering on the putative domains
- Creates a binary tree of clusters
- Each cluster is a **putative domain family**



- Machine learning & Scoring w.r.t. PfamA
- Choosing good families (intrinsic properties) – training/ disjoin to test
- Each family modelled by HMM, redefine **EV families**.
- Iteration (3 times from 100K to 25K)
- Jointing HMMs and voting for EV consensus family.

Method

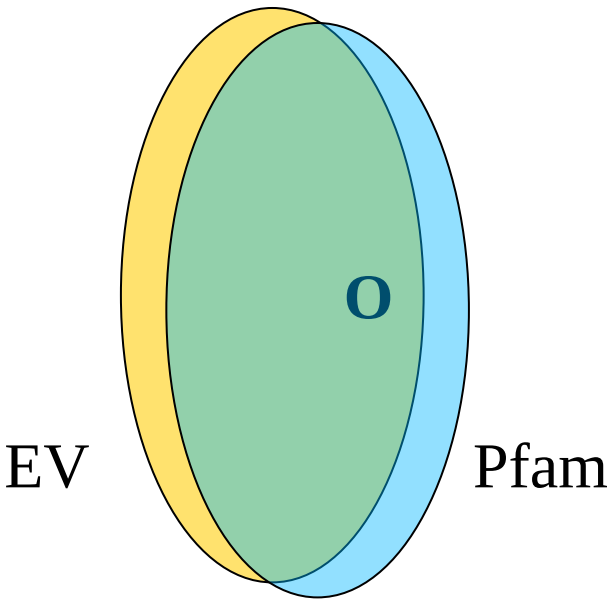
Quality & Evaluation

Comparing with **Pfam**

Pfam is a domain signature DB, manual curation, covers 62% aa, 7500 signatures

Accuracy – how well a typical EVEREST domain family scores w.r.t Pfam

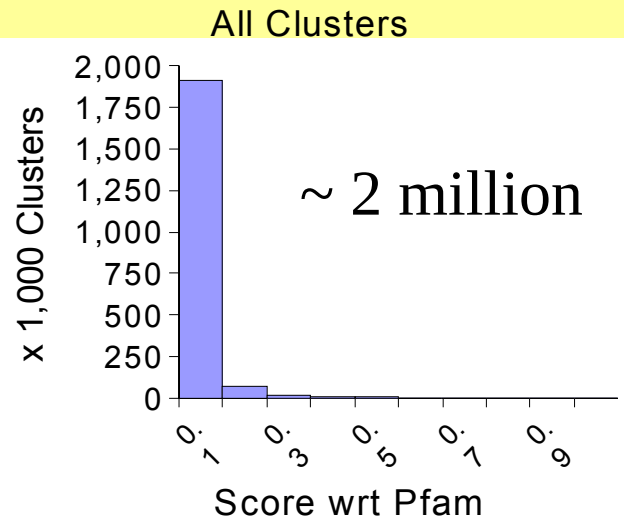
Size of the intersection over the size of the union
Scores range from 0 to 1.0 (Jaccard Score)

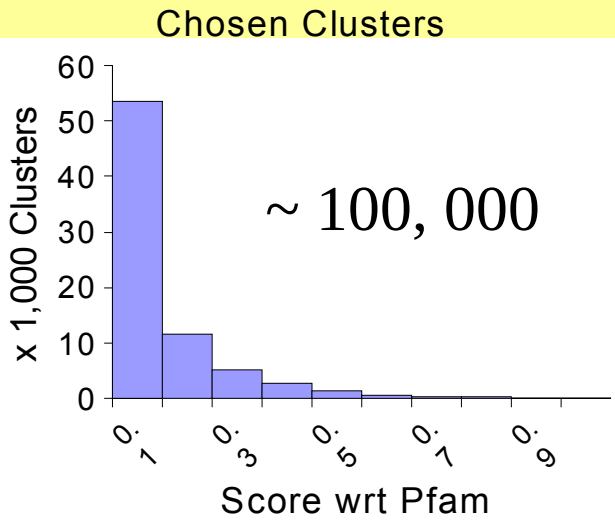


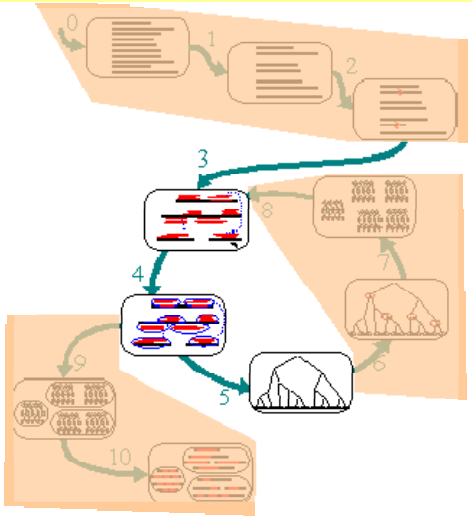
EV of 10 instances matches Pfam with 10 with only 9 are overlapping

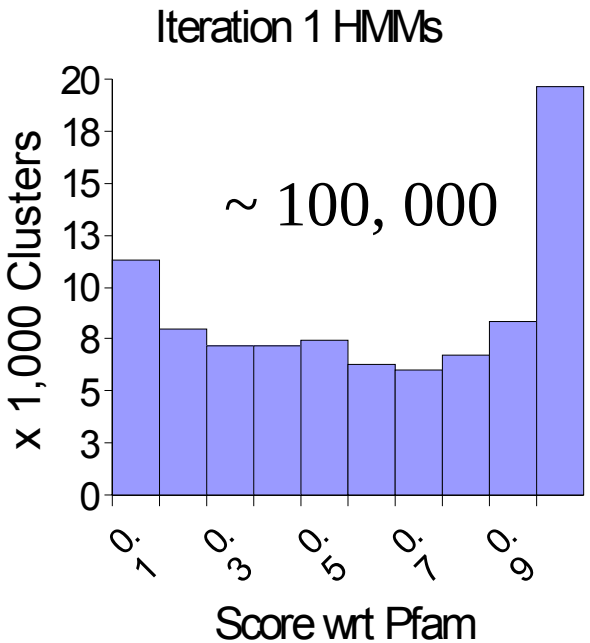
Score: 0.81

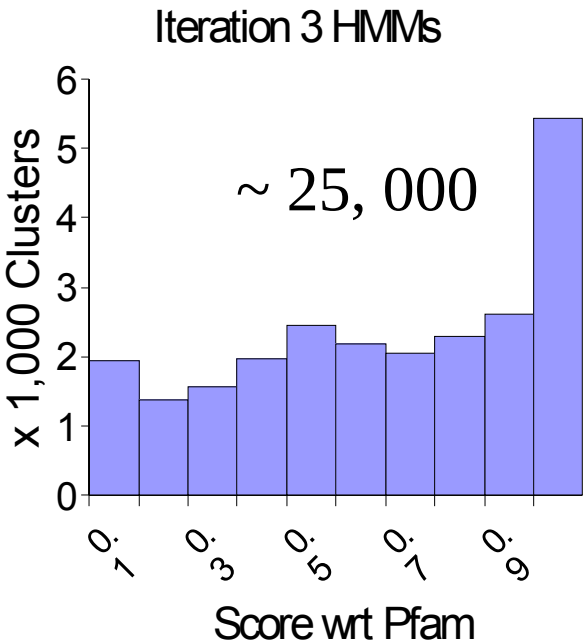
Getting Better (accuracy measure)

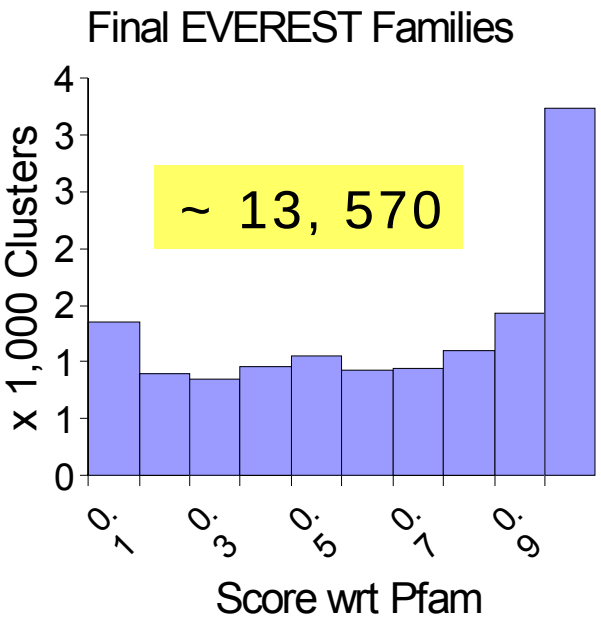








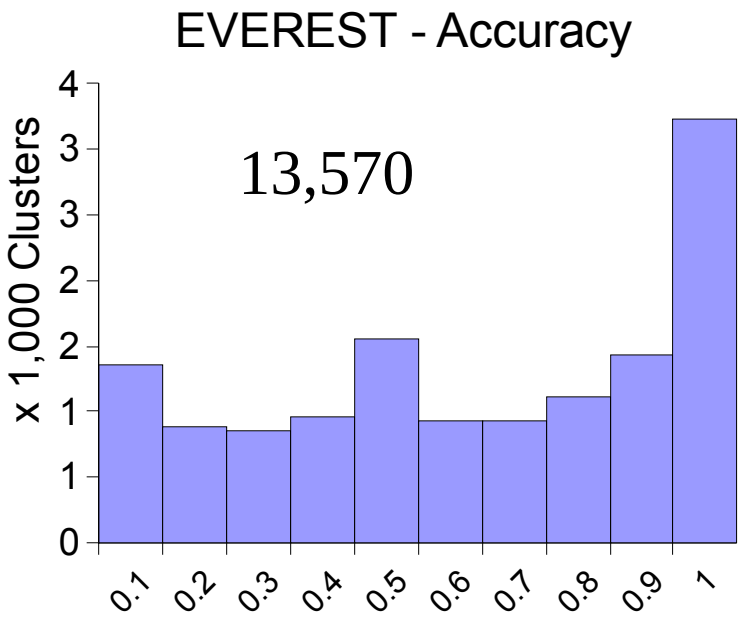


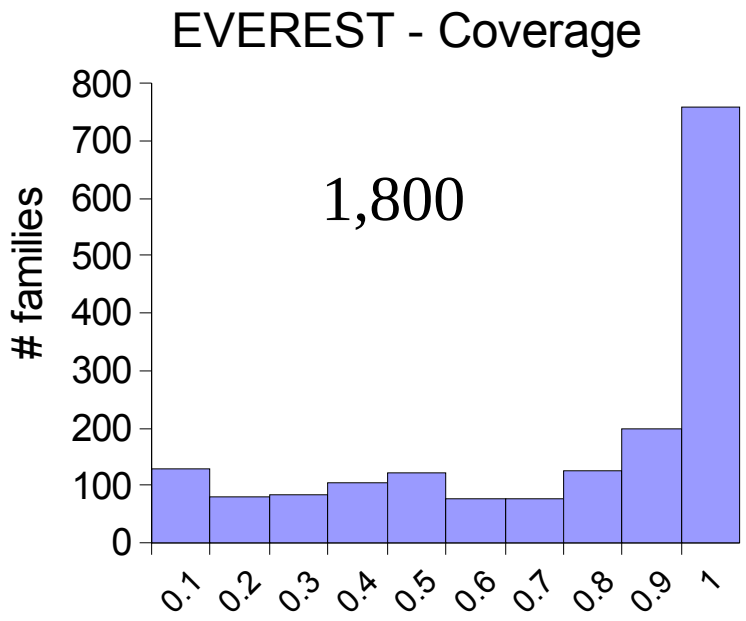


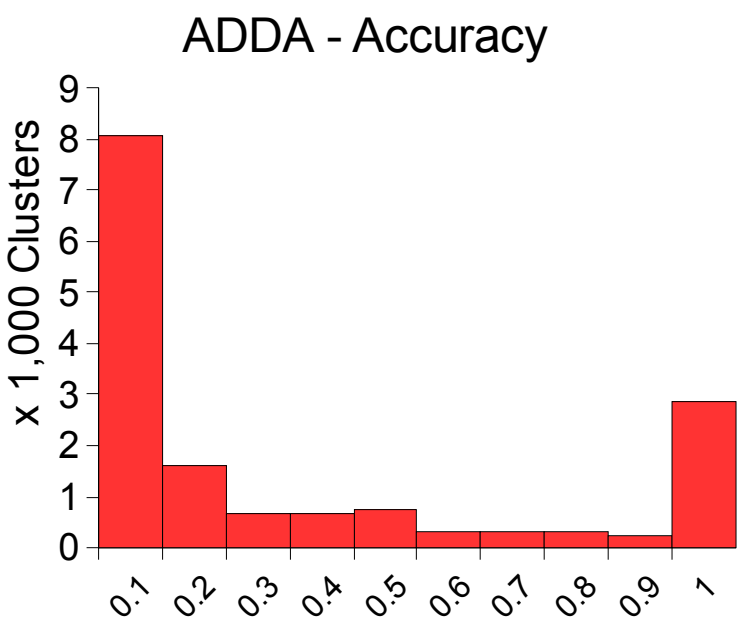
EVEREST – Evaluation vs Reference

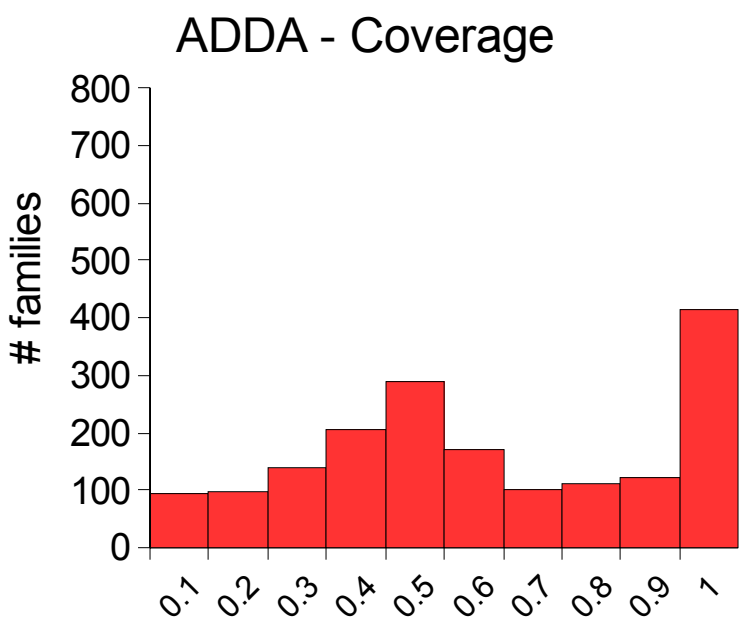
- EVEREST is evaluated against reference sets of known families (Pfam, SCOP, CATH)
- Score of EVERSET family w.r.t. Intersecting reference family:
  - size of intersection / size of union
  - **Accuracy**
    - Each EVEREST family scored vs. best matching reference
    - Look at score profile across EVEREST families
    - Ignore EVEREST families unknown to reference set
  - **Coverage**
    - Each reference family scored vs. best matching EVEREST
    - Look at score profile across interesting subsets of refrence set
    - Non-Trivial: family size>=5
    - Hetero: non-trivial + appearing in hetero-multi-domain proteins

Evaluation –wrt Pfam EVEREST & ADDA (Holm)

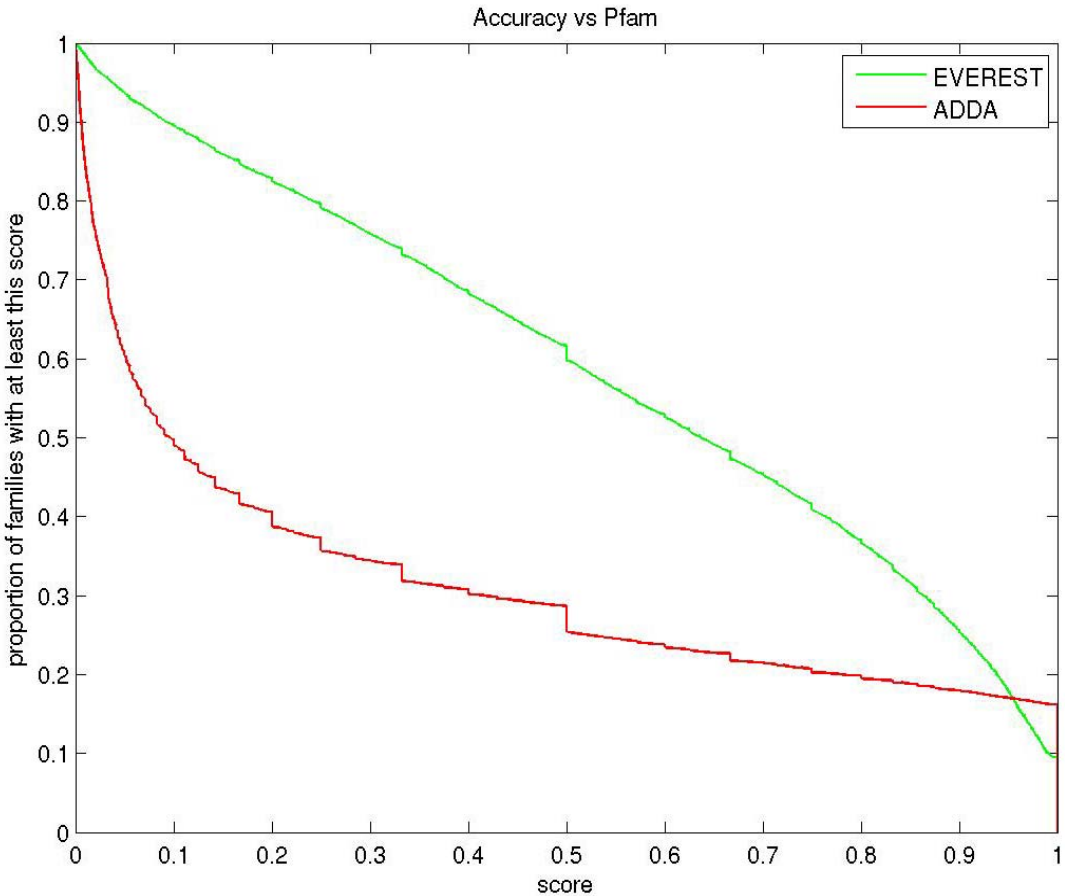


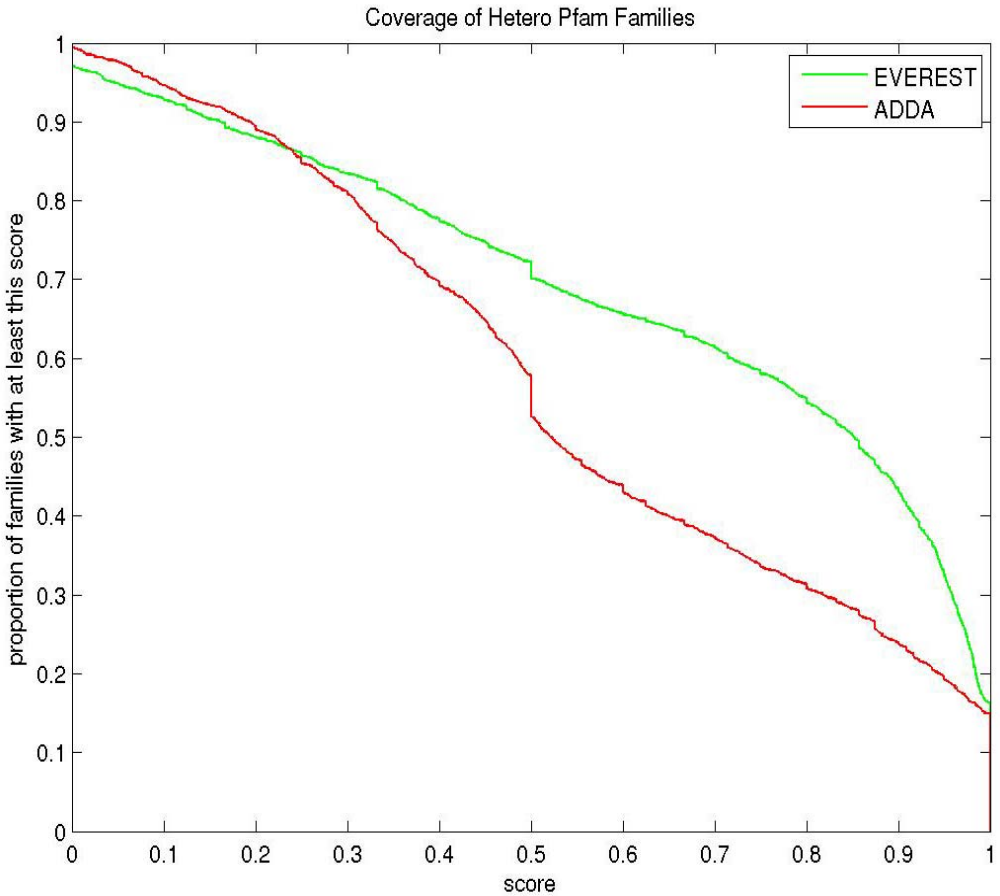






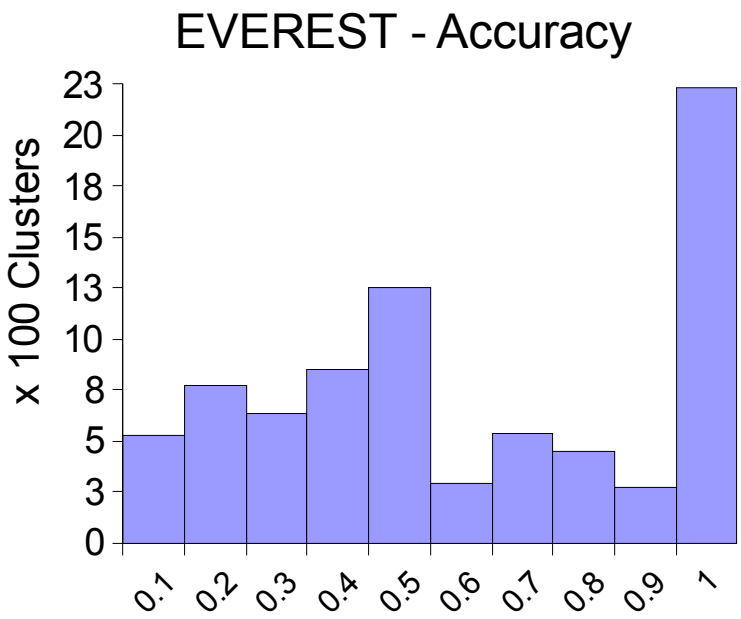
EVEREST & ADDA
Evaluation vs Pfam

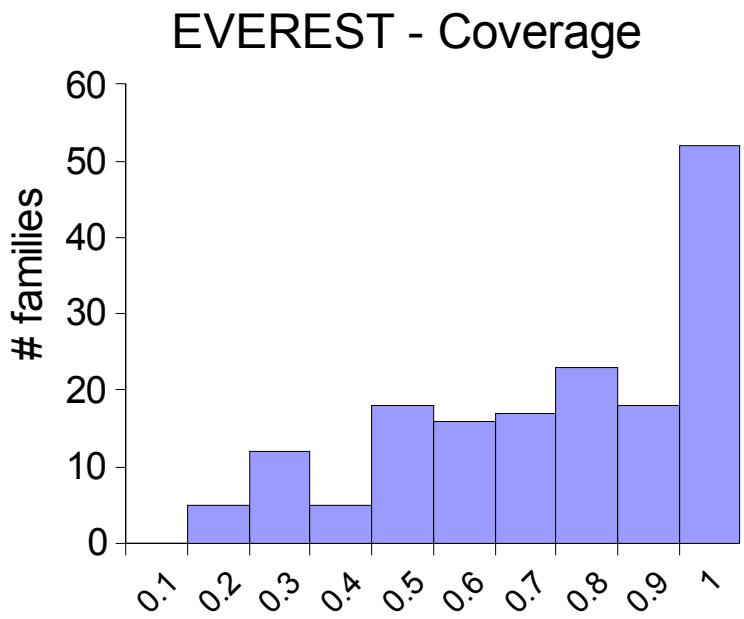




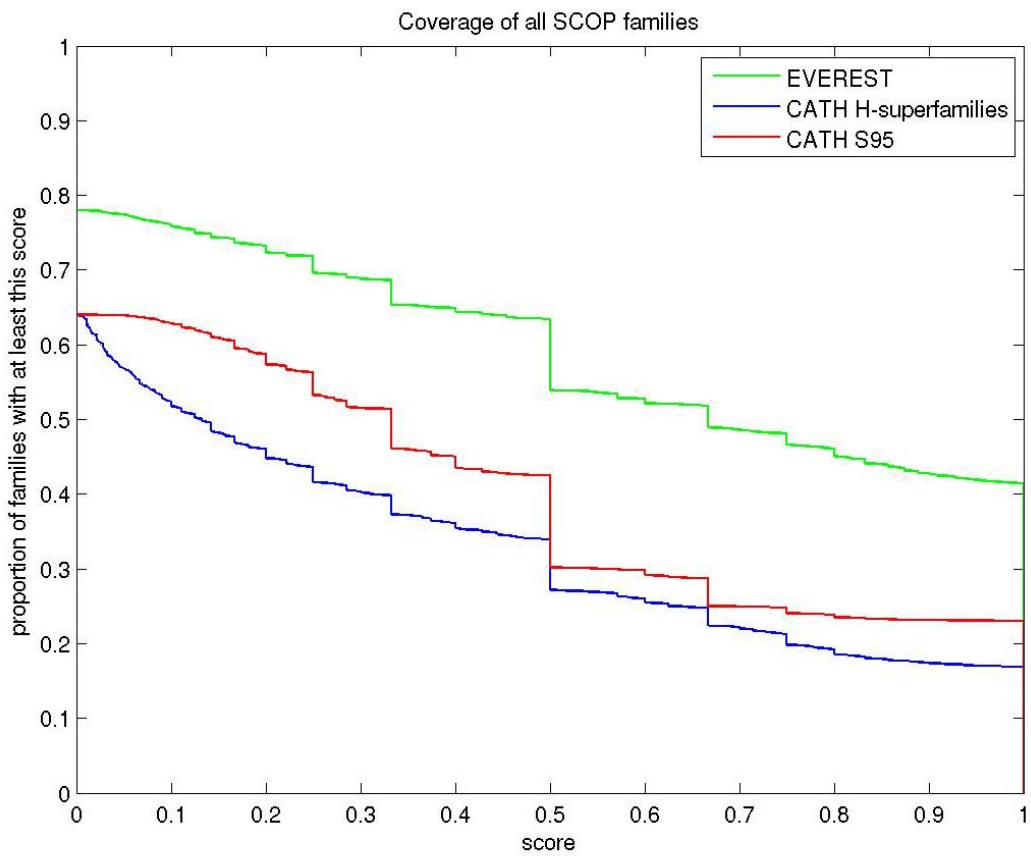
Hetero >5

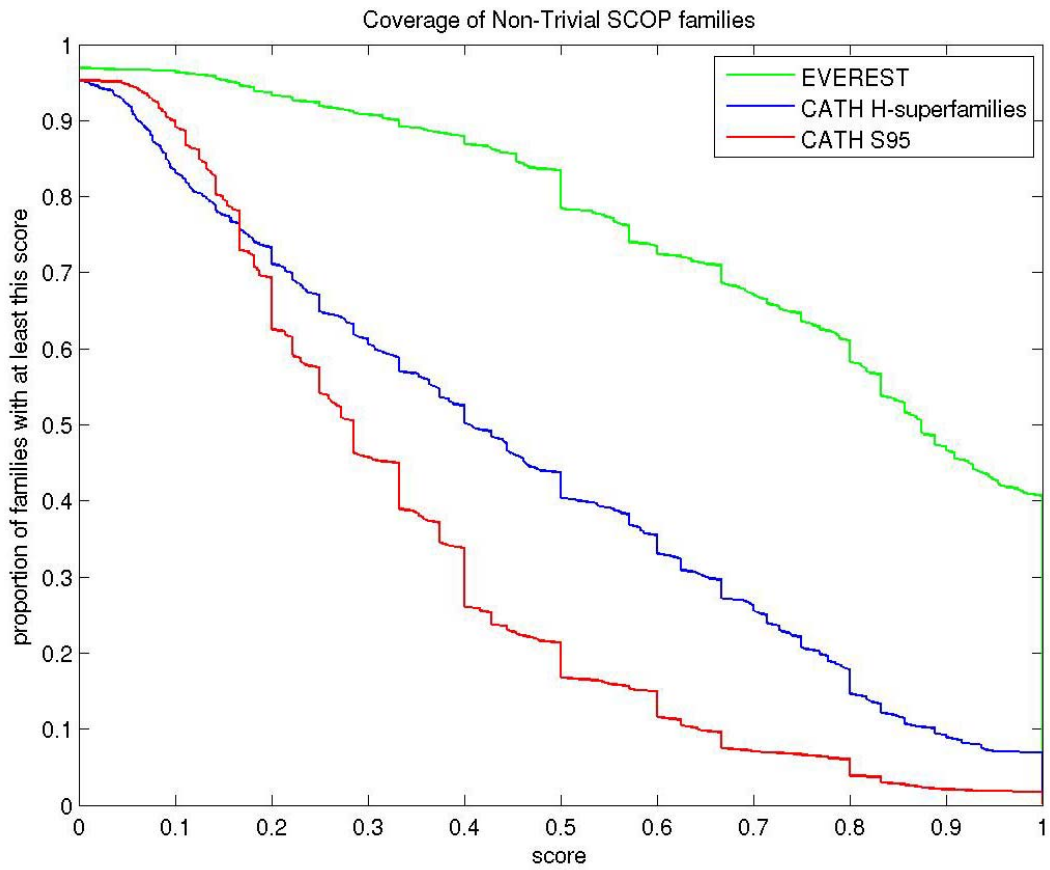
Evaluation – Compare w.r.t SCOP
manual classification of structural domains



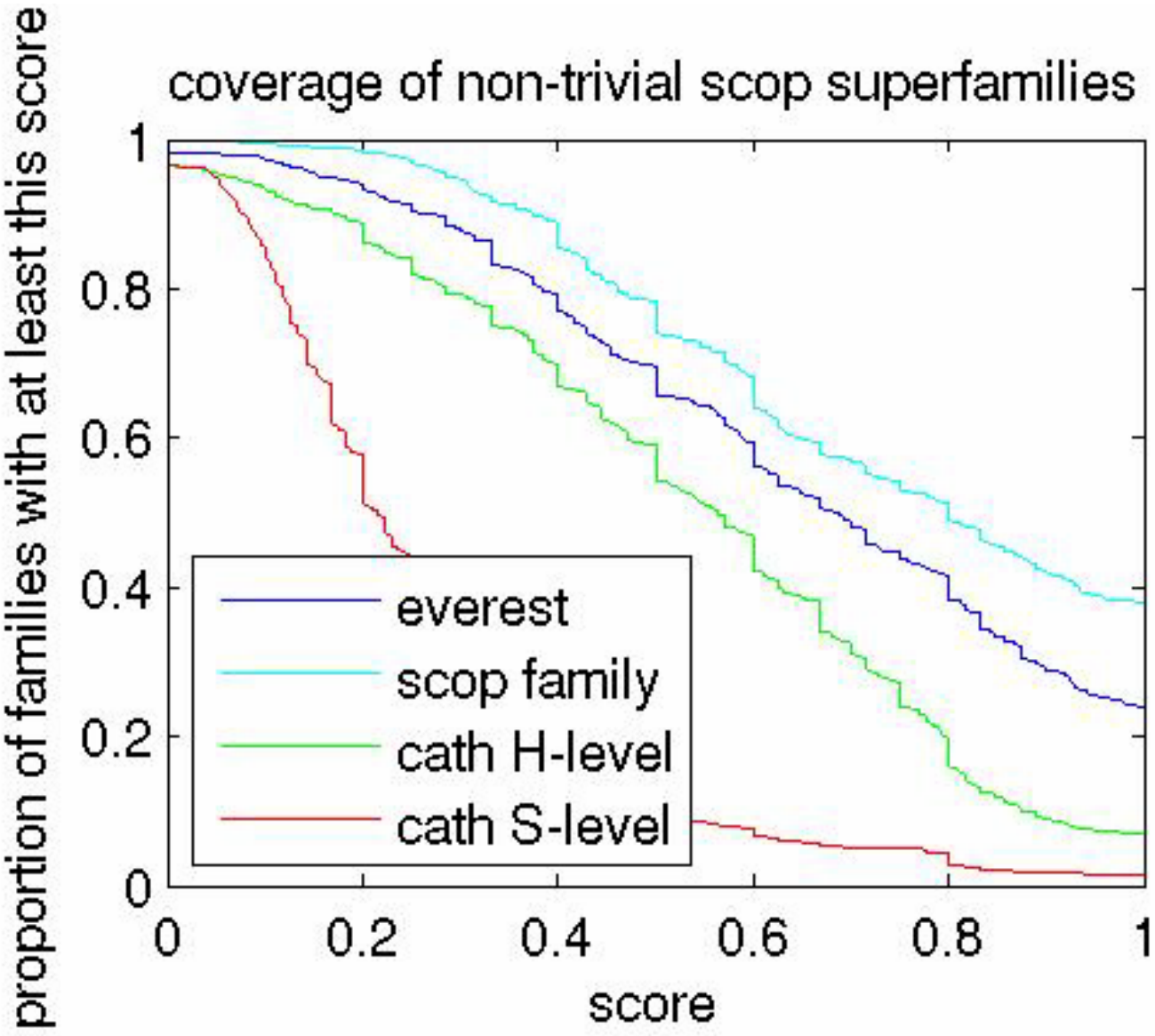


EVEREST – Evaluation vs SCOP (family) coverage





Evaluation – Compare wrt CATH / SCOP superfamily (coverage)



Overall Numbers
(for UniProt / SWP)

13,569 EV families were defined. Providing Joint HMMs.

Jointly cover **83% of the aa** in the SWP DB.

The average (median) size of an **EVEREST domain family** is 81 (41).

The average (median) **length of the domains** is 117 (76) aa.

Move to some examples (web based querying)

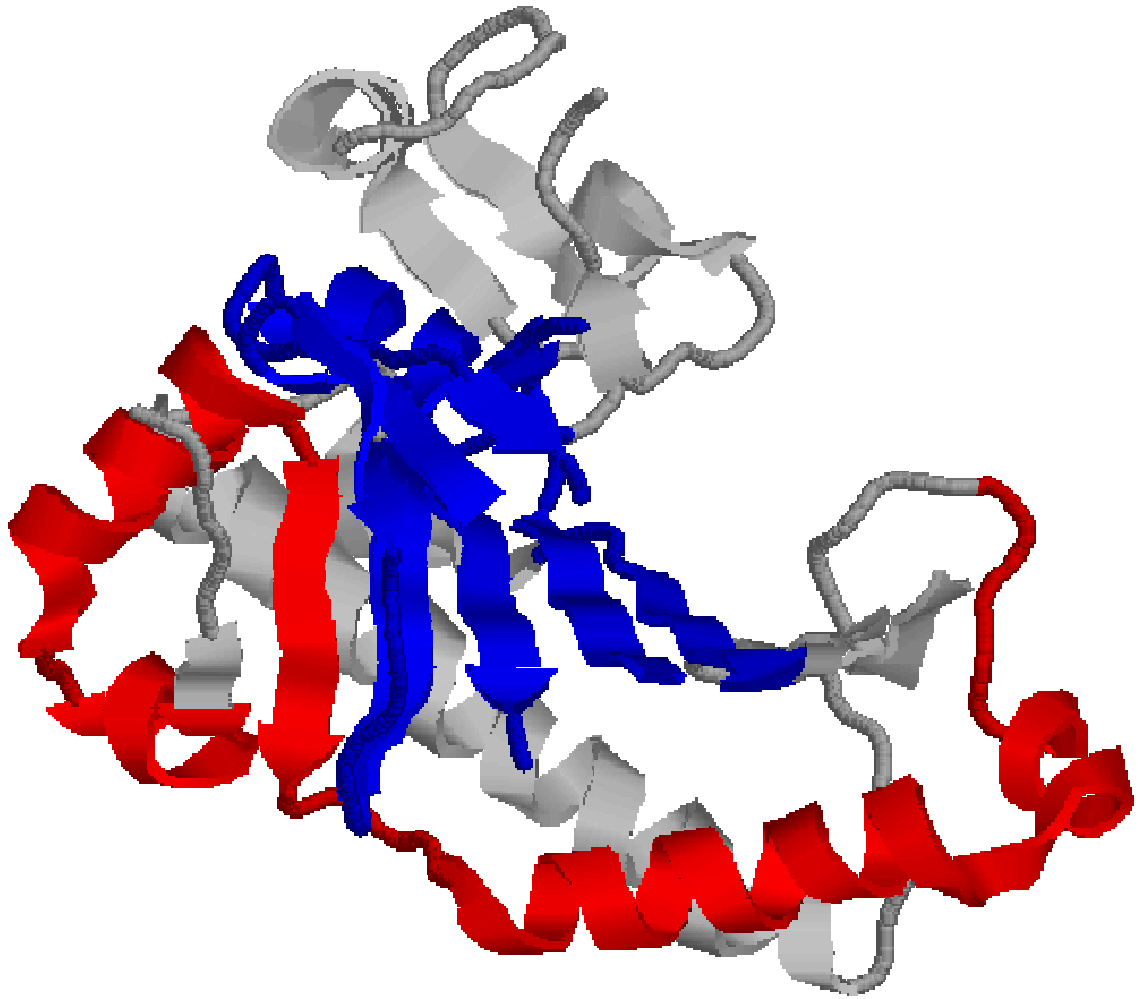
Examples:
New Functional Annotation

EVEREST family 1017
PF04673 (Polyketide synthesis cyclase)
PF04486 (SchA/CurD like protein)

- PF04486 has no known function
- Two of its members are known to be in gene clusters involved in the synthesis of polyketide-based spore pigments.
- Could these two families be considered one?

New Family (1)

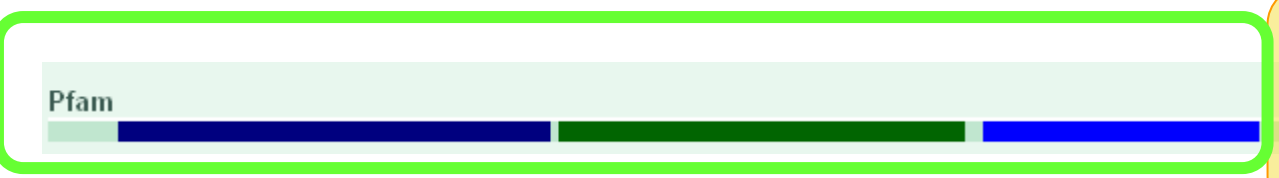
- EV02275 is unknown to Pfam
- 54 out of its 55 domains appear 90 positions N-terminal to PF03171 (2OG-Fe(II) oxygenase superfamily)
- Perhaps this is a new domain family?
- PDB 1UOG
  - **RED – EVEREST 2275**
  - **BLUE - PF03171**



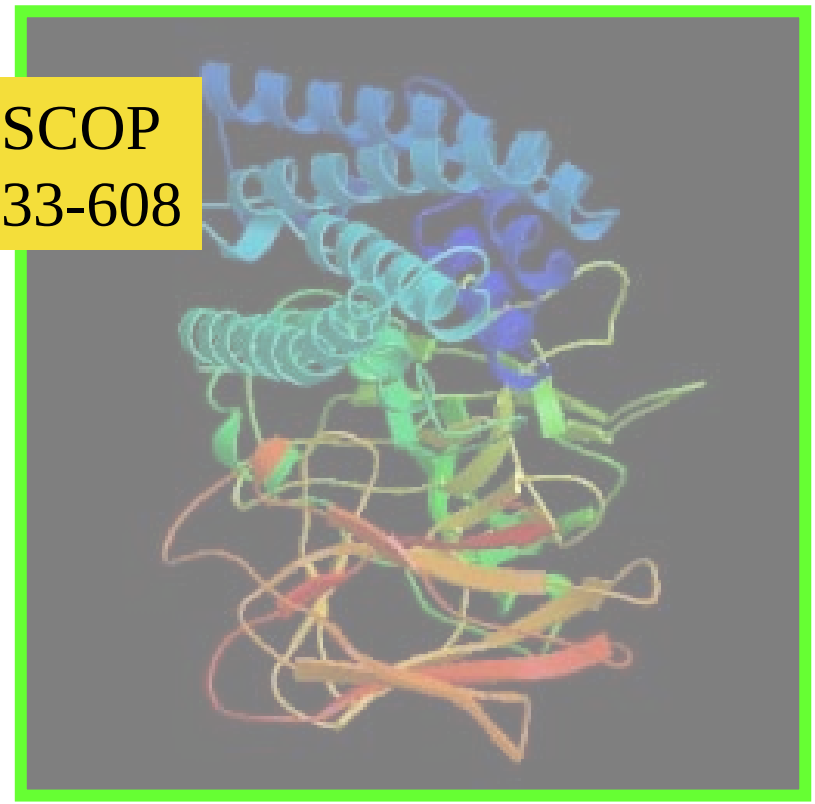
New domain family (2)

48 proteins – Pesticidial crystal protein cry5Aa
(Insecticidal delta-endotoxin CryVA(a) (Crystaline entomocidal protoxin)

EV covers the 48 proteins of PFAM (and SCOP / CATH) - perfectly



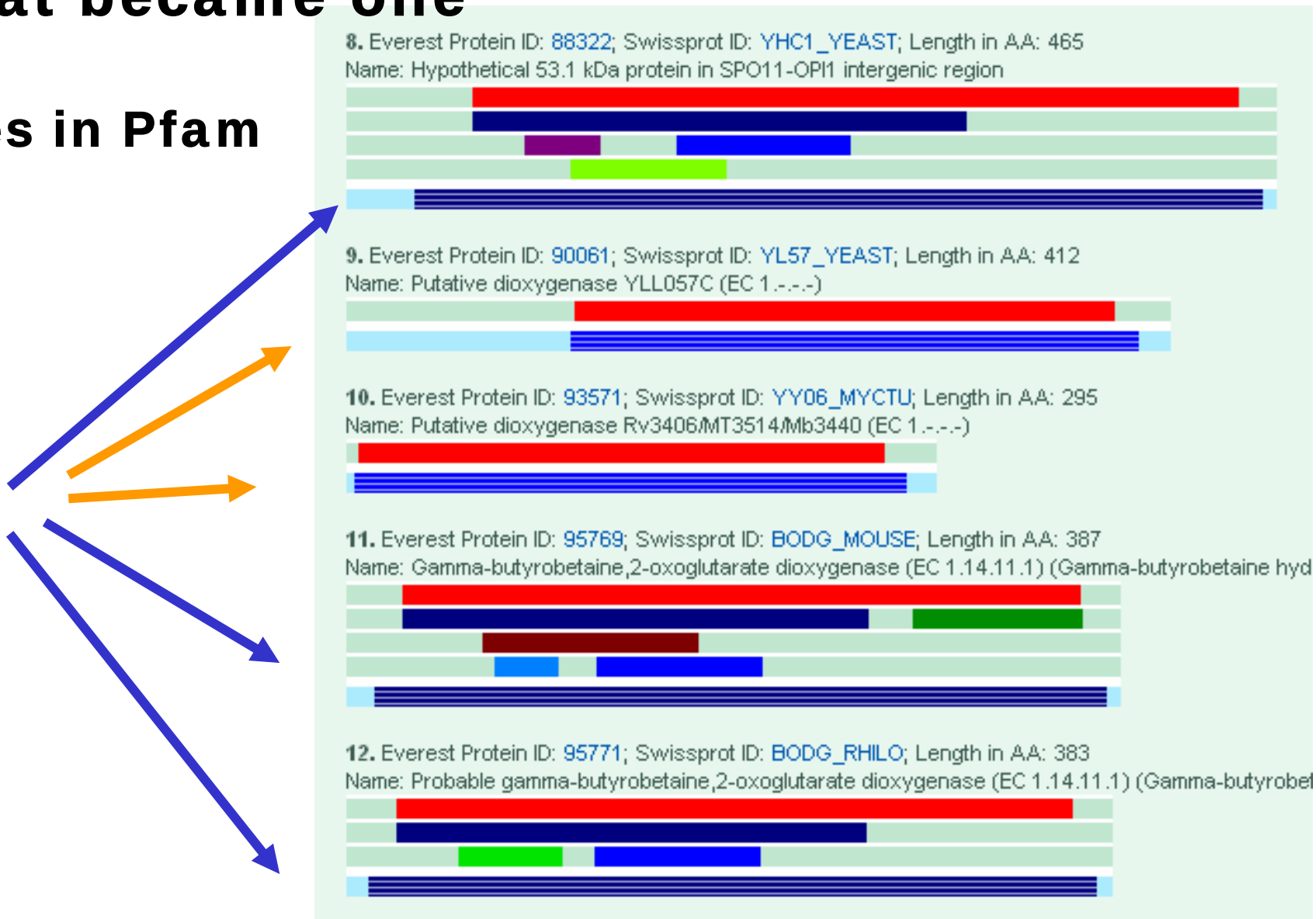
EVEREST



but another EV specifies the family – no OVERLAP and NO structure for this region (609-911)

Two that became one

Examples in Pfam CLANs

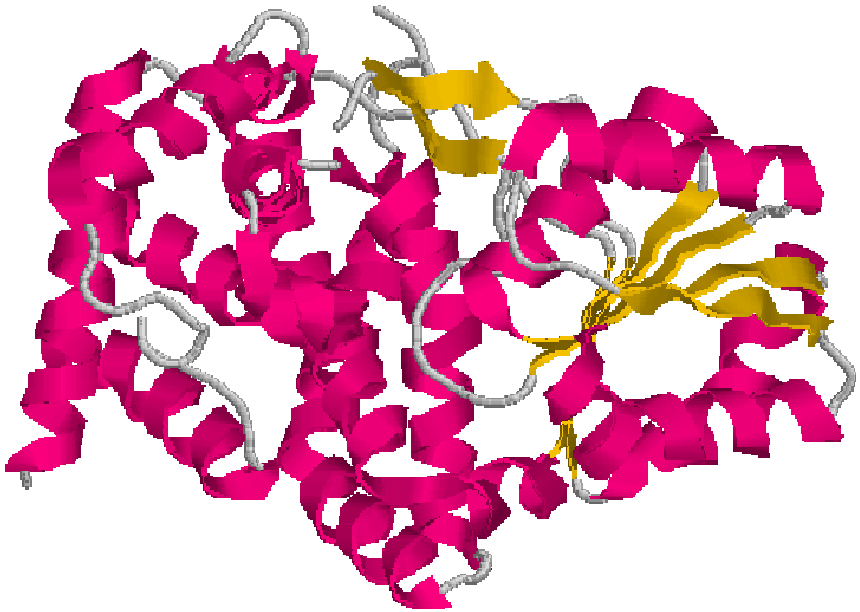


PFAM (OLD) Taurine catabolism dioxygenase TauD, TfdA family
Pfam (NEW) a composed entry: **TauD**

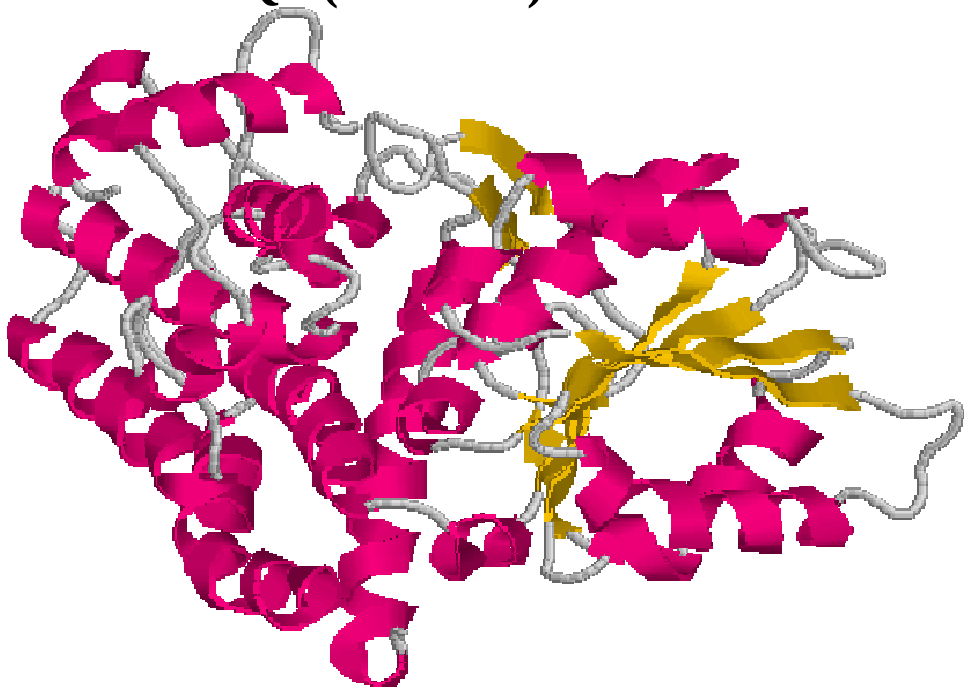
Superfamily

- EVEREST family **EV04463** fully covers both PF00465 (Iron-containing alcohol dehydrogenase) and PF01761 (3-dehydroquinate synthase).
- ENZYME: PF00465 is EC1.1-
- ENZYME: PF01761 is sometimes EC4.6 and sometimes EC1.1
- SCOP /CATH: Same superfamily/ Homology group

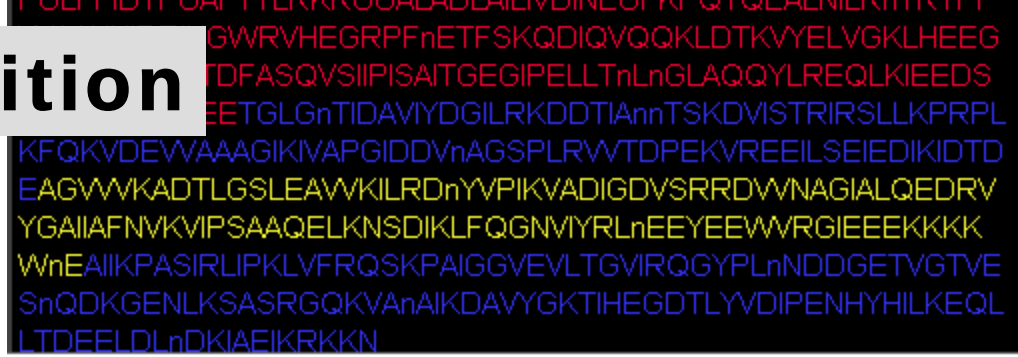
PDB 1JQA (PF00465)



PDB 1DQS (PF01761)

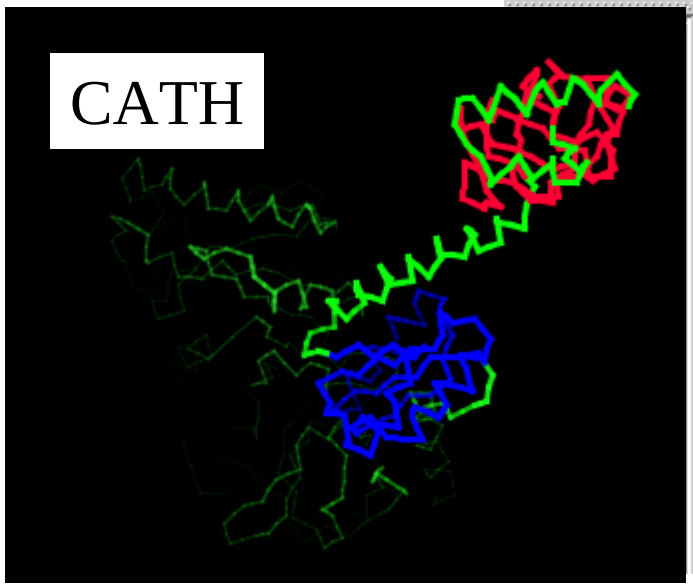


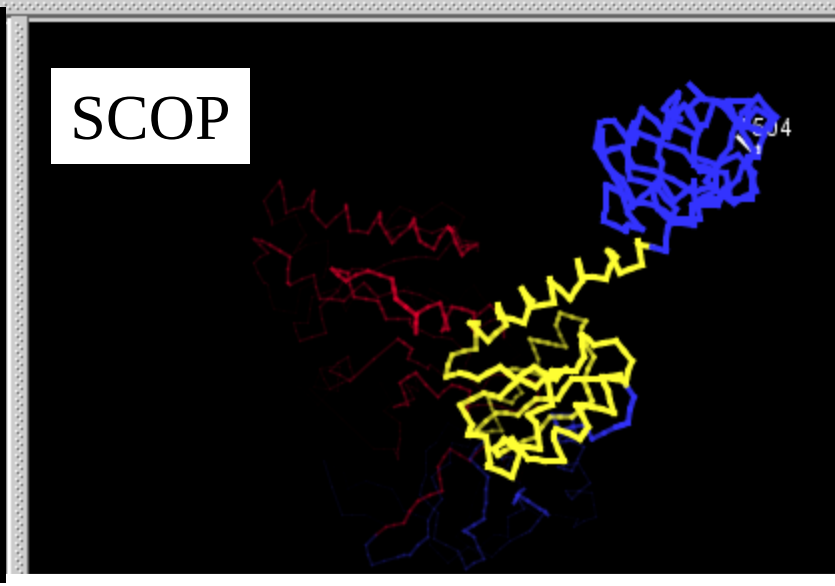
Alternative Family Definition

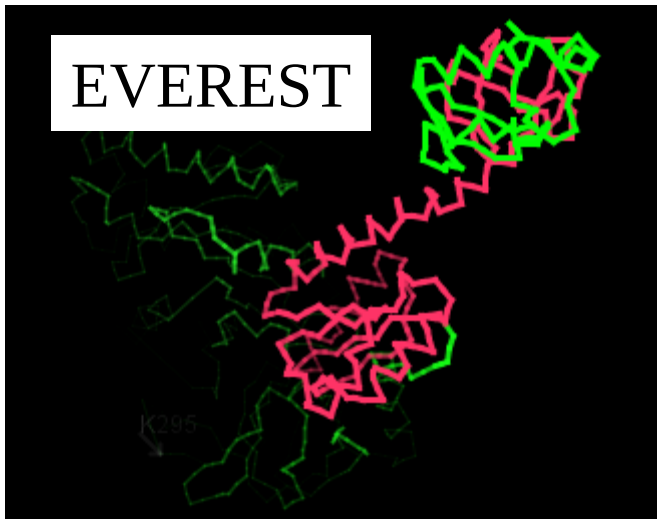


Elongation Factor

3 ‘domain family’ : All support same proteins







Half C-terminal
SCOP - two adjacent domains (yellow, blue)
CATH – two separated (blue, red) spacer (green)
EVEREST – one domain (pink)

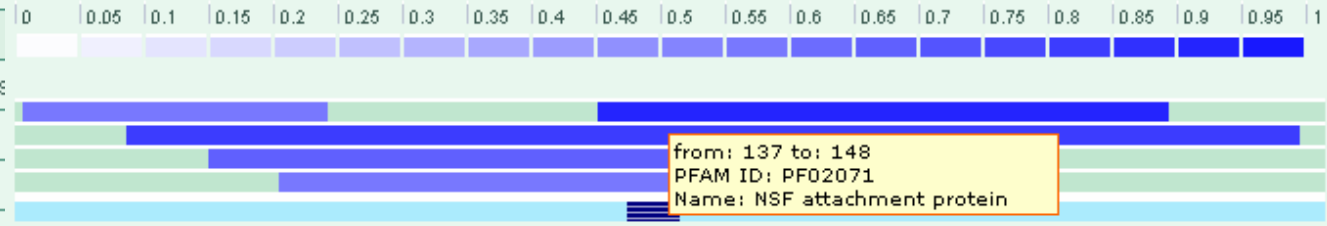
On the Web

Proteins search

Protein P-64671

Everest ID	P-64671
System	SwissProt 40.28
ID in Source	SC17_YEAST
Accession number	P32602
EMBL Protein-ID	
Protein name	Vesicular-fus
Length in amino acids	291
Theoretical pI	4.96
Molecular weight:	32757 Da
PDB	

Scores:



Sequence of prot

```
  1 MSDPVELLKRAEKKGVPS
    YKFEEAADLCVQAATIYR
 61 LKAADYQKKAGNEDEAGN
    GNSVNAVDSLENPIQIFT
121 FELGEILENDLHDYAKAI
    DQSVALSNKCFIKCADLK
181 YSKLIKSSMGNRLSQWSL
    AATDAVAARTLQEGQSED
241 KSLIDAVNEGDSEQLSEH
    WKITILNKIKESIQQQED
```

Keywords

Swissprot	Endoplasmic reticulum, Golgi stack, Protein transport, Transport
InterPro accession number	IPR000744
GO	**GO cellular component:** Golgi apparatus, Cell, Cellular_component, Cytoplasm, Endoplasmic reticulum, Intracellular **GO molecular function:** Intracellular transporter, Molecular_function, Protein transporter, Transporter **GO biological process:** Biological_process, Cell growth and/or maintenance, Intracellular protein transport, Protein transport, Transport

NCBI Taxonomy

```
SUPERKINGDOM - eukaryota
  |_ KINGDOM - fungi
    |_ PHYLUM - ascomycota
      |_ SUBPHYLUM - saccharomycotina
        |_ CLASS - saccharomycetes
          |_ ORDER - saccharomycetales
            |_ FAMILY - saccharomycetaceae
              |_ GENUS - saccharomyces
                |_ SPECIES - saccharomyces cerevisiae
```

Evaluate any reference domain resources

Display settings:
Choose sequence databases
Choose domain family systems

Display Settings:

Swiss-Prot	PDB	SCOP	CATH	Pfam	Pfam Clans
Non Redundant	Non Redundant	Family level	S35 level	On	Off

Apply

View EVEREST Family EV00014

Number of proteins in PDB: 4 (representative: 2).
Number of proteins in Swiss-Prot: 16 (representativ
Total proteins: 20 (representative: 15).

Download HMMs:

	1	2	3
Global-Local HMM:			
Global-Global HMM:			

- List of domains of EV00014 in tabular form.
- Scoring of EV00014 by families from other systems.
- Scoring of families from other systems by EV00014.

Family page header:
General statistics
Download of HMMs
Links to list of domains and to evaluation pages

1. Everest Protein ID: 11000603; PDB ID: 1SU0 Length in AA: 159
Name: NifU like protein IscU

EVEREST

No.	PFAM ID	PFAM Name	Family ID	TP	FP	FN	Score	Percent Known to Pfam
1	PF02071	NSF attachment protein	5755	44	10	0	0.81	4% 96%
2	PF02071	NSF attachment protein	12525	26	0	18	0.59	37% 63%

Legend:
Families Appearances

Family ID	Count	Count total
10369	9	25
12204	9	25
1875	9	25
10564	9	25
2328	5	17
6667	4	8
2118	3	7
11462	1	1
11047	1	10
11787	1	3

Total 25 proteins. Representative: 9 proteins

Downloads: 1
Global-local HMM:
Global-global HMM:

View list of all proteins | View list of representative proteins only | View PFAM intersections with family 10369

1. Everest Protein ID: 65047; Swissprot ID: SEC1_YEAST; Length in AA: 724
Name: Protein transport protein SEC1

2. Everest Protein ID: 65772; Swissprot ID: SLP1_CAEEL; Length in AA: 576
Name: Protein slp-1

3. Everest Protein ID: 65806; Swissprot ID: SLY1_YEAST; Length in AA: 666
Name: SLY1 protein

4. Everest Protein ID: 67547; Swissprot ID: STB2_CANFA; Length in AA: 593
Name: Syntaxin binding protein 2 (Unc-18 homolog 2) (Unc-18B) (Unc18-2)

Name: Vacuolar protein sorting-associated protein 45

6. Everest Protein ID: 114200; Swissprot ID: VP33_YEAST; Length in AA: 691
Name: Vacuolar protein sorting 33 (SLP1 protein)

7. Everest Protein ID: 114204; Swissprot ID: VP3A_HUMAN; Length in AA: 596
Name: Vacuolar protein sorting 33A (hVPS33A)

8. Everest Protein ID: 114207; Swissprot ID: VP3B_HUMAN; Length in AA: 617
Name: Vacuolar protein sorting 33B (hVPS33B)

9. Everest Protein ID: 114211; Swissprot ID: VP45_MOUSE; Length in AA: 570
Name: Vacuolar protein sorting-associated protein 45 (mVps45)

Is there any added value for The overlapping EV families?

EV10564 /100% - perfect match but 220 aa not 640 aa

EV01875/ 87% cover / 3 new

Legend: **EVEREST Family** Neighbors

	Neighbor	Type	Level	Forward	Backward
	EV00014	Same		(15/15) (15|0)	(15/15) (15|0)
	EV08449	Sub	1	(14/15) (14|0)	(14/21) (14|0)
	EV07109	Sub		(11/15) (11|1)	(11/22) (11|1)
	EV05061	C-Term		(5/15) (5|0)	(5/46) (5|0)
	EV01411	C-Term		(4/15) (4|1)	(4/18) (4|1)
	EV04683	Super		(1/15) (1|1)	(1/30) (1|1)
	EV09954	Same		(0/15) (0|4)	(0/8) (0|4)
	EV09838	Same		(0/15) (0|1)	(0/3) (0|1)
	Other families				

Legend: **SCOP Family** Neighbors

	Neighbor	Type	Level	Forward	Backward
	d.224.1.2	Same		(2/2) (2|0)	(2/2) (2|0)
	Other families				

Legend: **PFAM Family** Neighbors

	Neighbor	Type	Level	Forward	Backward
	PF01592	Same		(12/13) (10|1)	(12/14) (10|1)
	PF01106	C-Term		(5/13) (4|0)	(5/15) (4|0)
	PF04324	C-Term		(3/13) (3|1)	(3/20) (3|1)
	Other families				

Download HMMs:

	1	2	3
Global-Local HMM:			
Global-Global HMM:			

- List of domains of EV00014 in tabular form.
- Scoring of EV00014 by families from other systems.

Family color code legend:
Current family always in red
Relationship of current family to other families

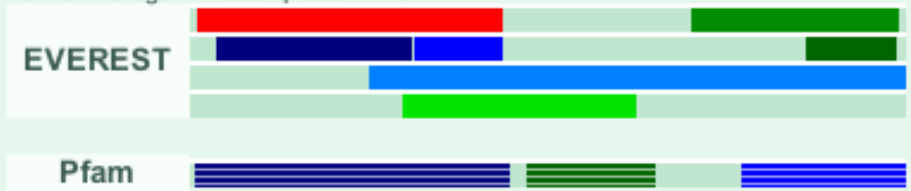
Type refers to relationship between boundaries:

same = similar boundaries
subdomain
superdomain
C-terminal neighbor
N-terminal neighbor

Forward =" how many of the member of the current family participate in the relationship

Backward =" how many of the member of the other family participate in the relationship

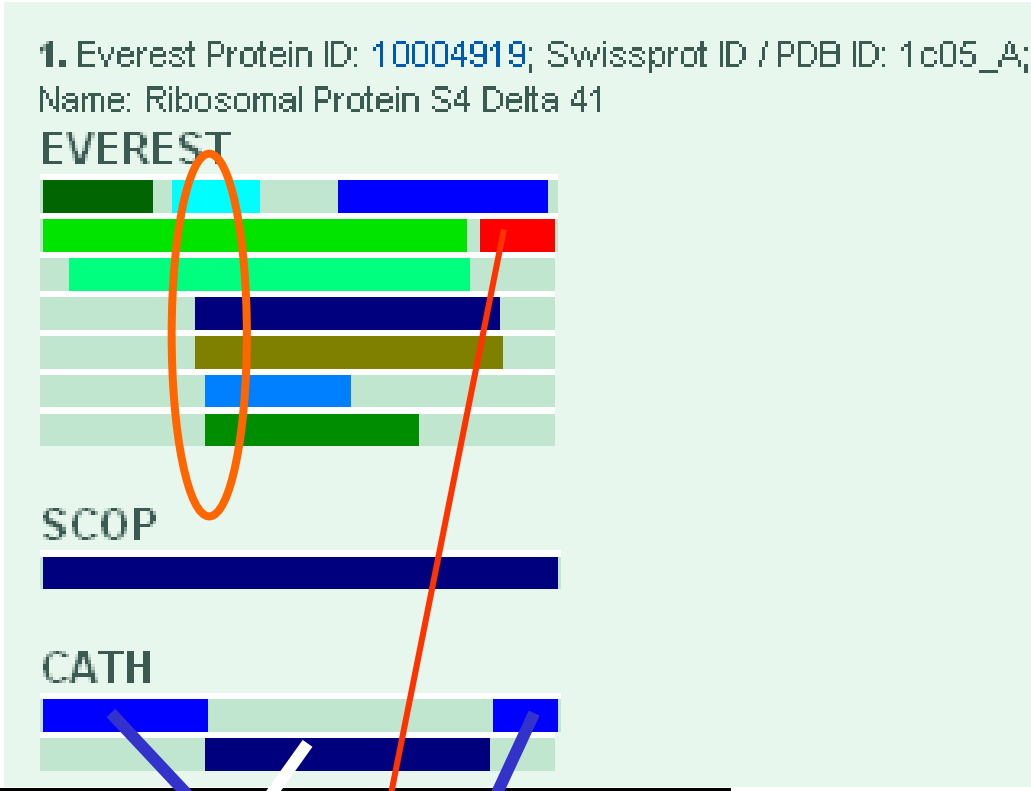
3. Everest Protein ID: 45068; Swiss-Prot ID: NIFU_ANAAZ; Length in AA: 300
Name: Nitrogen fixation protein nifU

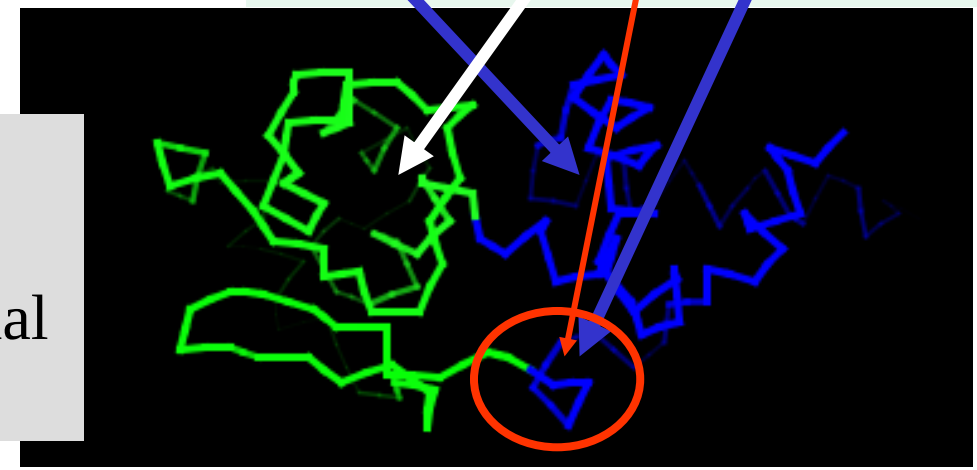


4. Everest Protein ID: 45071; Swiss-Prot ID: NIFU_AQUAE; Length in AA: 157
Name: NifU-like protein

Next Phase:

- Improving EVEREST web
- **Evaluation** of ALL used resources
- Phylogenetic View
- Enrich queries (according to reference Resource)
- Names for EVxxxx
- **Paste** your protein
- Domain **boundaries**





79 proteins

30S ribosomal protein S4

Summary:

- We provide an automated framework for identification and classification of new protein domains

  - recovering 60% of difficult known Pfam families.
  - Suggests new families for 8% (with > 51% fidelity)
  - For 20% we suggest a new view on domain families

- Manual inspection of families scoring low w.r.t. Pfam suggested that many of those are valid families.

- Enabling inspection of EVEREST families and additional resources in **http://www.everest.cs.huji.ac.il**

Annotation & Inference
New genomes, New functions

EVEREST

Having Function

Experiments
Literature
Expert view

Automatic
(no pre-knowledge)

Partition to ‘domains’
(no transitivity)

Robustness
(evaluate w.r.t others)

No Function

New genomes
No similarity
No evidence

Annotation & Inference
New genomes, New functions

Domain families by EVEREST
Automatic identification of Protein Domain
Performance and analysis w.r.t to other resources

New Annotation by Inference
A method for inference – testing on a new genome-the BEE

New Function to Disserted Proteins
High level functionality – story of the toxin like proteins

Honey Bee
The brain & complex neuronal behavior

C Elegans (worm)		19,000
Miniat. Wasp		10,000
Drosophila (fruit fly)		14,000
Apis (honey bee)		10,000
Homo Sapiens		25,000

The number of neurons or genes is not indicative for the brain and behavior complexity.

The makeup of a social behaving insect

ProtoBee: Goal

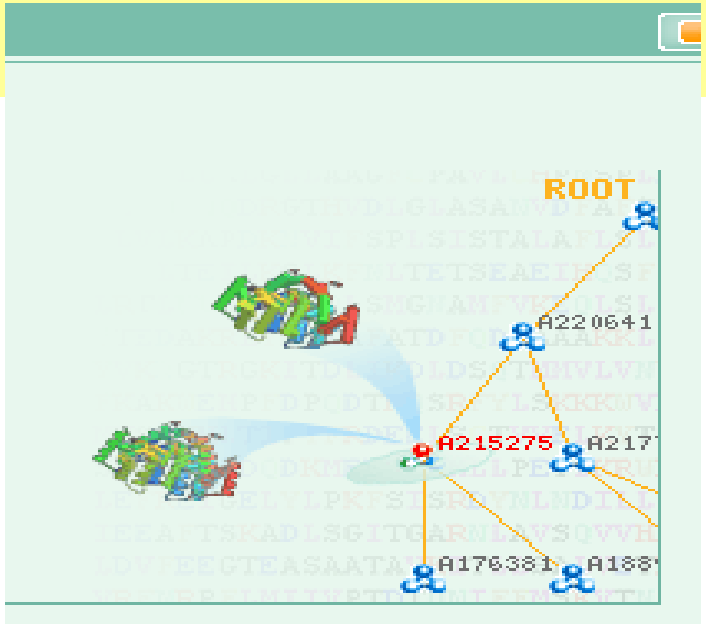
ProtoBee.cs.huji.ac.il

Honey bee genome recently sequenced: ~200 MB
(by HGSC at Baylor College of Medicine)

10,157 predicted ORFs

- Produce a hierarchical (functional) organization of the bee proteome
- **Annotate the bee sequences**
- Systematically find putative instances of
  - Bee gene-loss events
  - Bee-specific paralogs
  - Bee-specific functionality
  - Mis-predicted genes (FN/FP)

ProtoNet classifications
The Principles: A reminder



- Unsupervised
- Only sequence information as input
- All proteins involved (incl. hypothetical..)
- Family definition is hierarchical
- Only based on statistical significance of the similarity score
- Clustering process after ALL mutual 'distance' information is computed (Blast of All against All for 120 K proteins, E=100)

Evaluation vs InterPro, GO etc

Pfam, Prosite, SMART, PRINTS, SCOP, CATH...

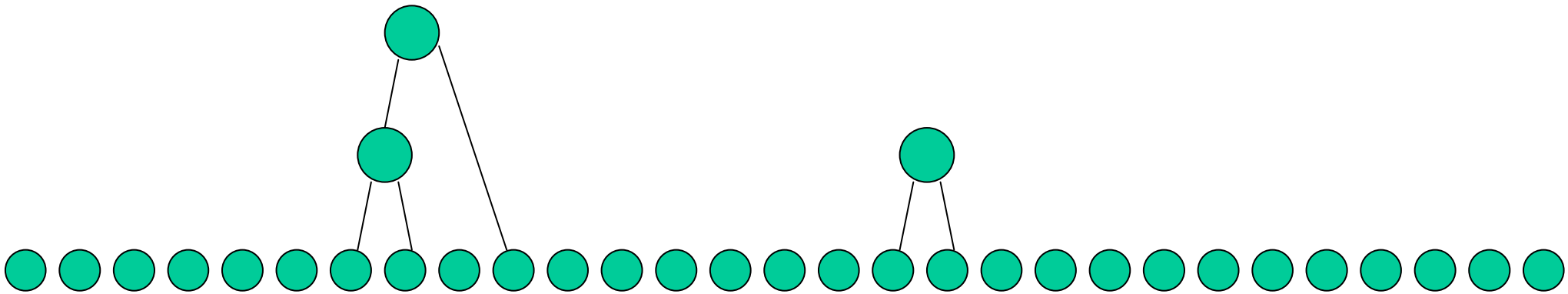
www.protonet.cs.huji.ac.il

Clustering Method

First, each protein is considered a singleton (a cluster of its own).

Clustering Method

- Next, we iteratively merge the pairs of clusters
- We choose to merge the ‘most similar’ pair of clusters.



Clustering Method

The clustering process gradually generates a tree of clusters

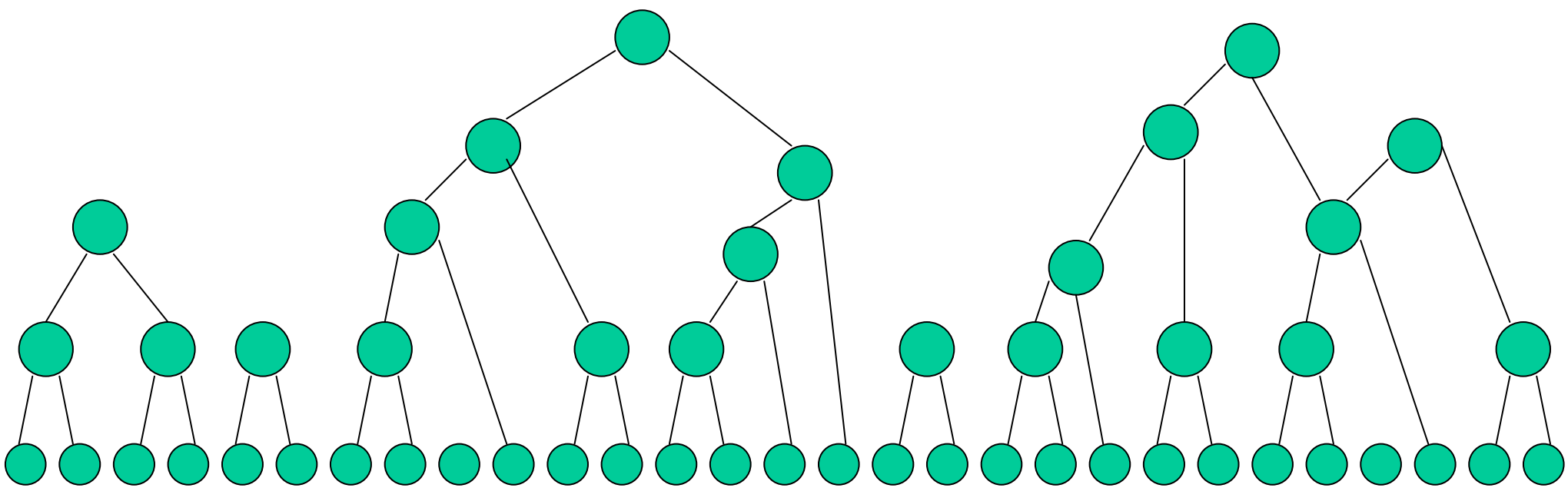
Merging Scores

$$\frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\sqrt[n]{x_1 x_2 \ldots x_n}$$

$$\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

Pruning:
Compact the tree to 12% of its size without
Reduction in performance (w.r.t. InterPro)



quality..

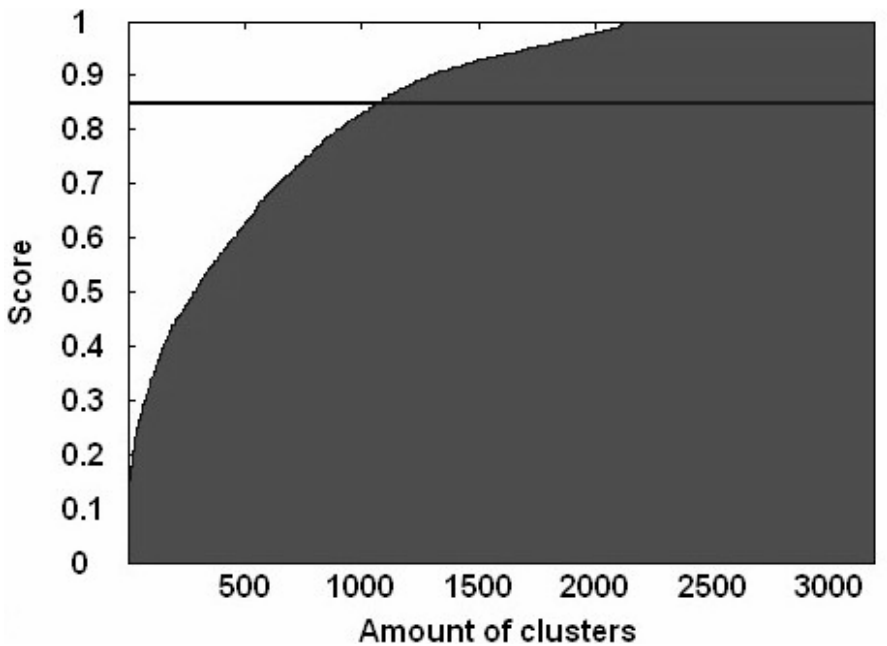
ProtoNet Hierarchical organization

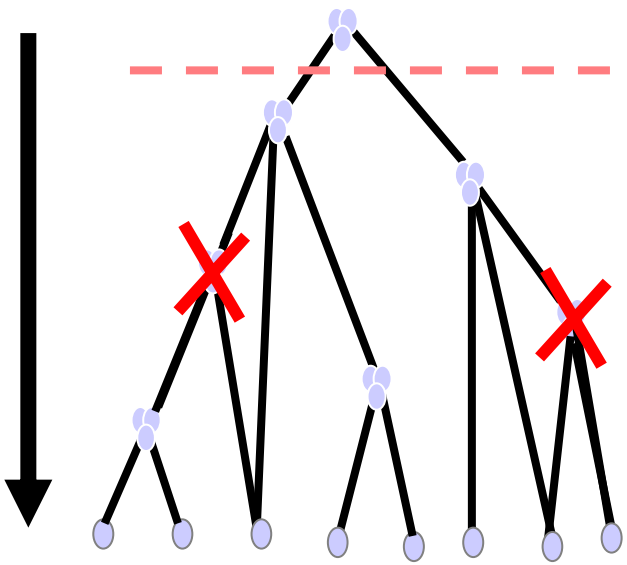
Protein database:

SwissProt ~133,000 proteins –

Testing the 'Matching Score' for InterPro (combining all high – quality domain based / structure base / knowledge based)

$$score(C,S) = \frac{|C \cap S|}{|C \cup S|}$$





Annotation Inference for proteins in clusters

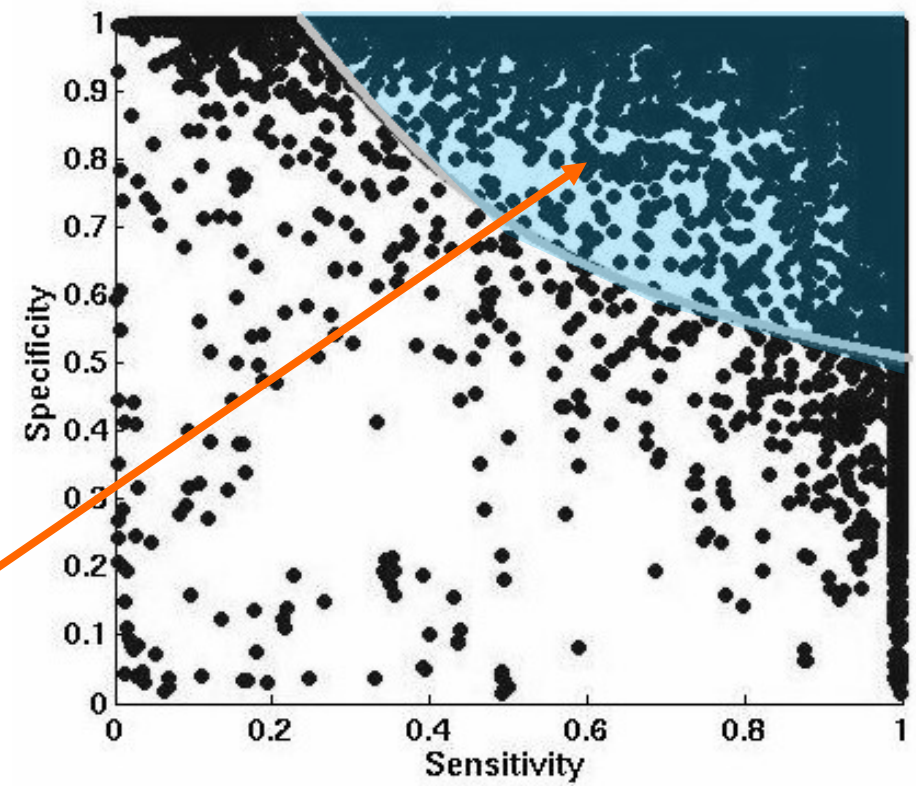
C- cluster C ; K - keyword

Annotation Score AS (C,K) = specificity^2 x sensitivity = 0.25

$$\left(\frac{TP}{TP+FP}\right)^2 \times \frac{TP}{TP+FN}$$

TP is the proteins in C that have the keyword K
FN is the proteins not in C that have the keyword K
FP is proteins in C that do not have the keyword K.

ProtoNet > 20 pr

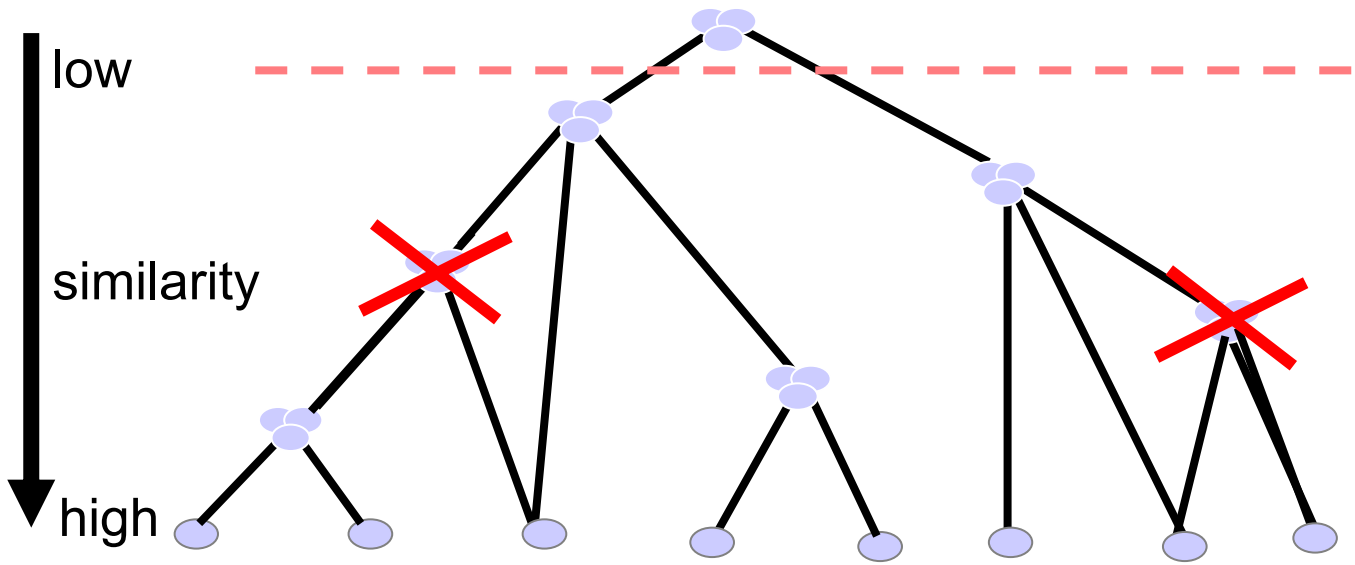


The high-confidence annotation threshold

Method for the Bee
Hierarchical organization

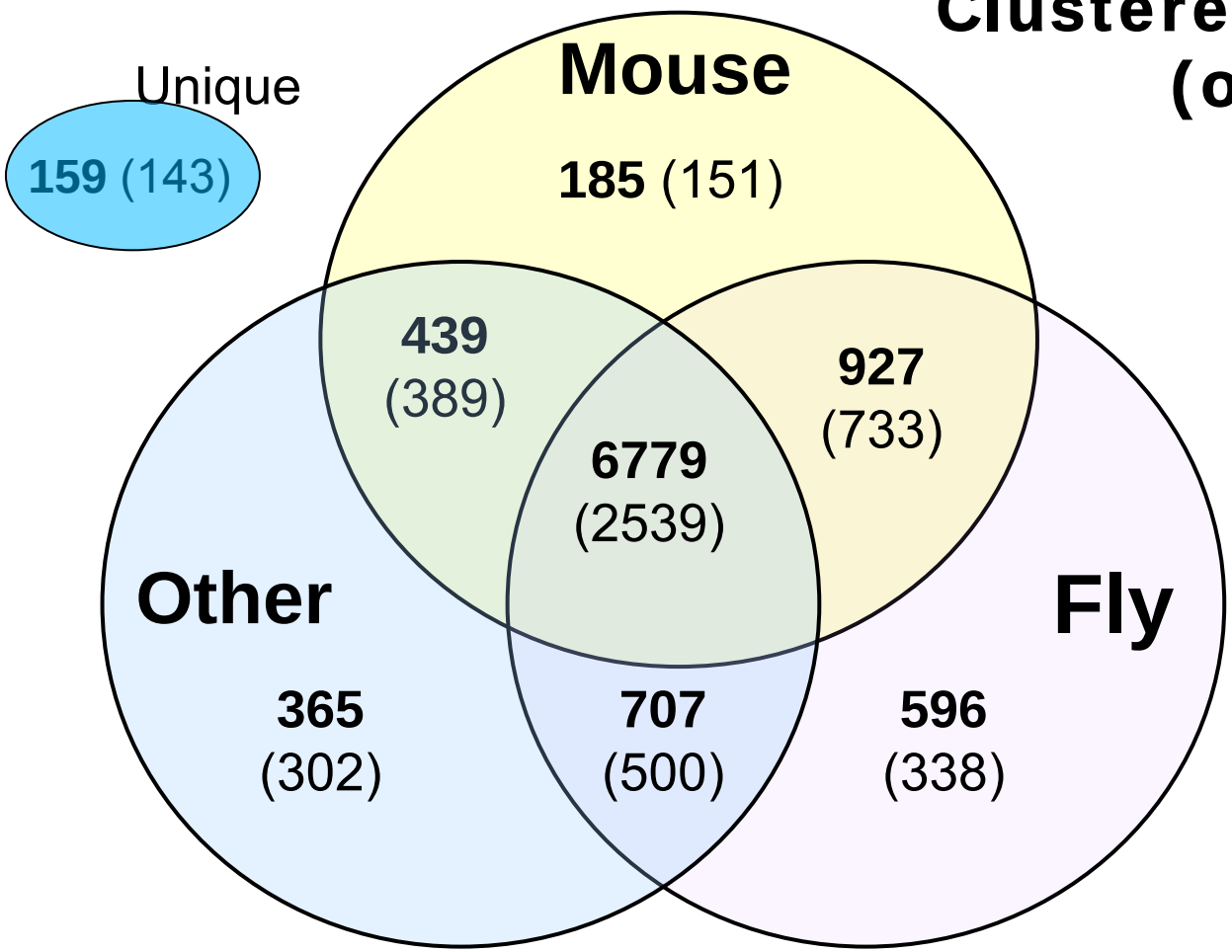
- Protein database (200,000 pr)
  - Predicted **bee** protein set: 10,157 pr
  - SwissProt (without bee) – ~133,000 proteins.
  - **Drosophila** proteome (insect) – 20,730 pr.
  - **mouse** proteome (UniProt) – 35,199 pr.

- All vs all BLAST
- Clustering
- Tree chopping
- Tree pruning



ProtoBee: results

Clustered into 5095 families (out of 18,500)



Bee annotation inference
high confidence

For each cluster, **calculate its annotations**. Each annotation is required to:

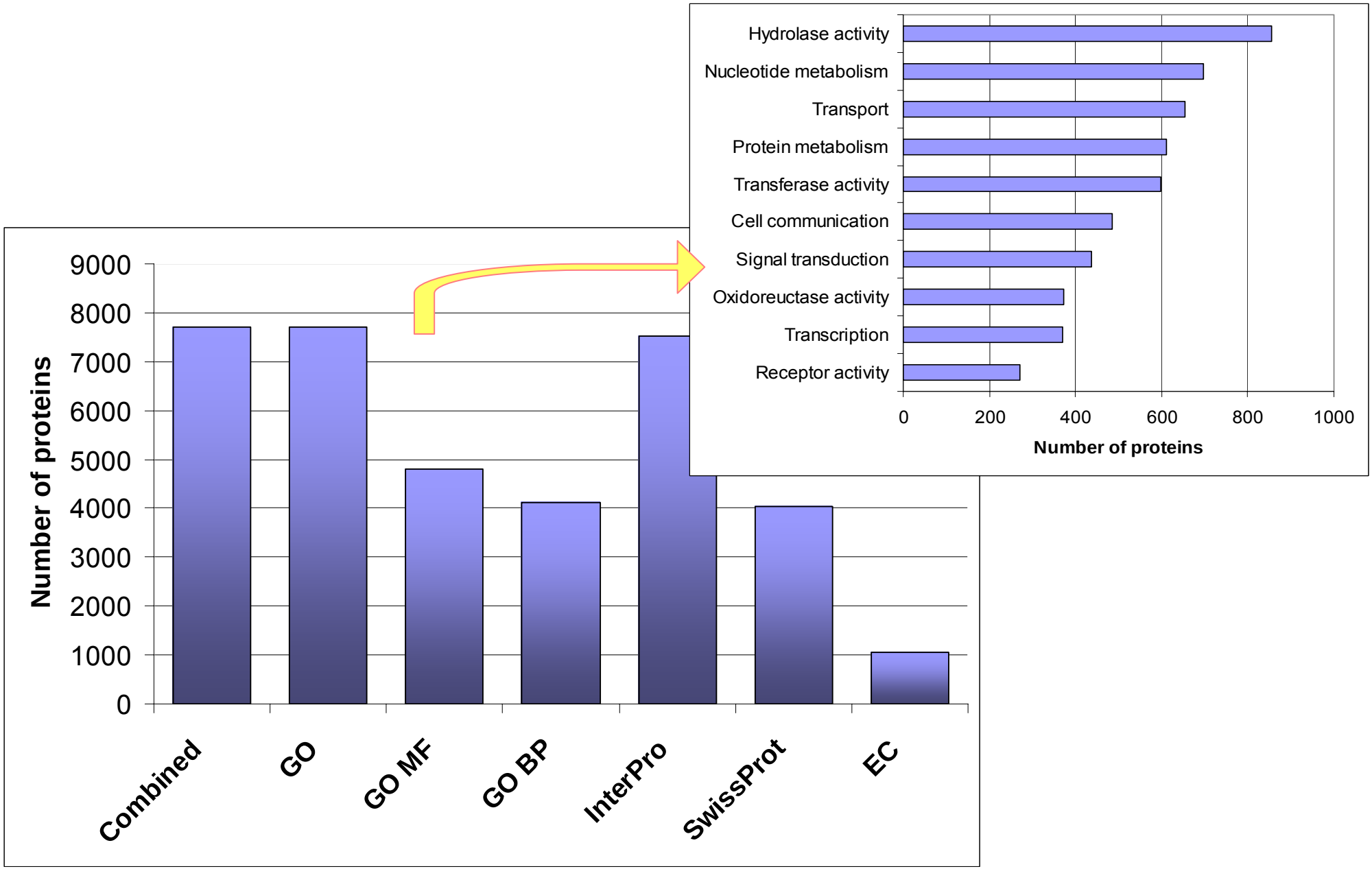
(a) be assigned > **75%** of the proteins in the cluster

(b) achieve **p-value <= 0.001** (hypergeometric distribution).

Only clusters with **> 5 proteins** are considered

For each bee protein, assign to it the annotations of its cluster and all parents.

Annotation summary



How good is this method?

Pros (assuming negligible transitivity):

- Any kind of external information source can be used for annotation.
- "Robustness" reduces chance of false positives.
- Potentially links biological properties to localized sequence features.

Cons:

- Incorrect transitivity due to multiple domains.
- Not as sensitive/specific as motif-based methods.

Results overview

Clusters organized into **18,936** trees (roots).

5095 roots contain **bee** proteins.

Annotation: 70% of proteins are annotated (InterProScan covers ~72-78%).

Interesting biological information on the evolution of the bee relative to other insects (different talk)

Annotation & Inference
New genomes, New functions

Having Function

Experiments
Literature
Expert view

ProtoBee

Annotation Score
(high confidence)

Clusters leading to
Retesting ORFs

No Function

New genomes
No similarity
No evidence