

Protein family sub-classification

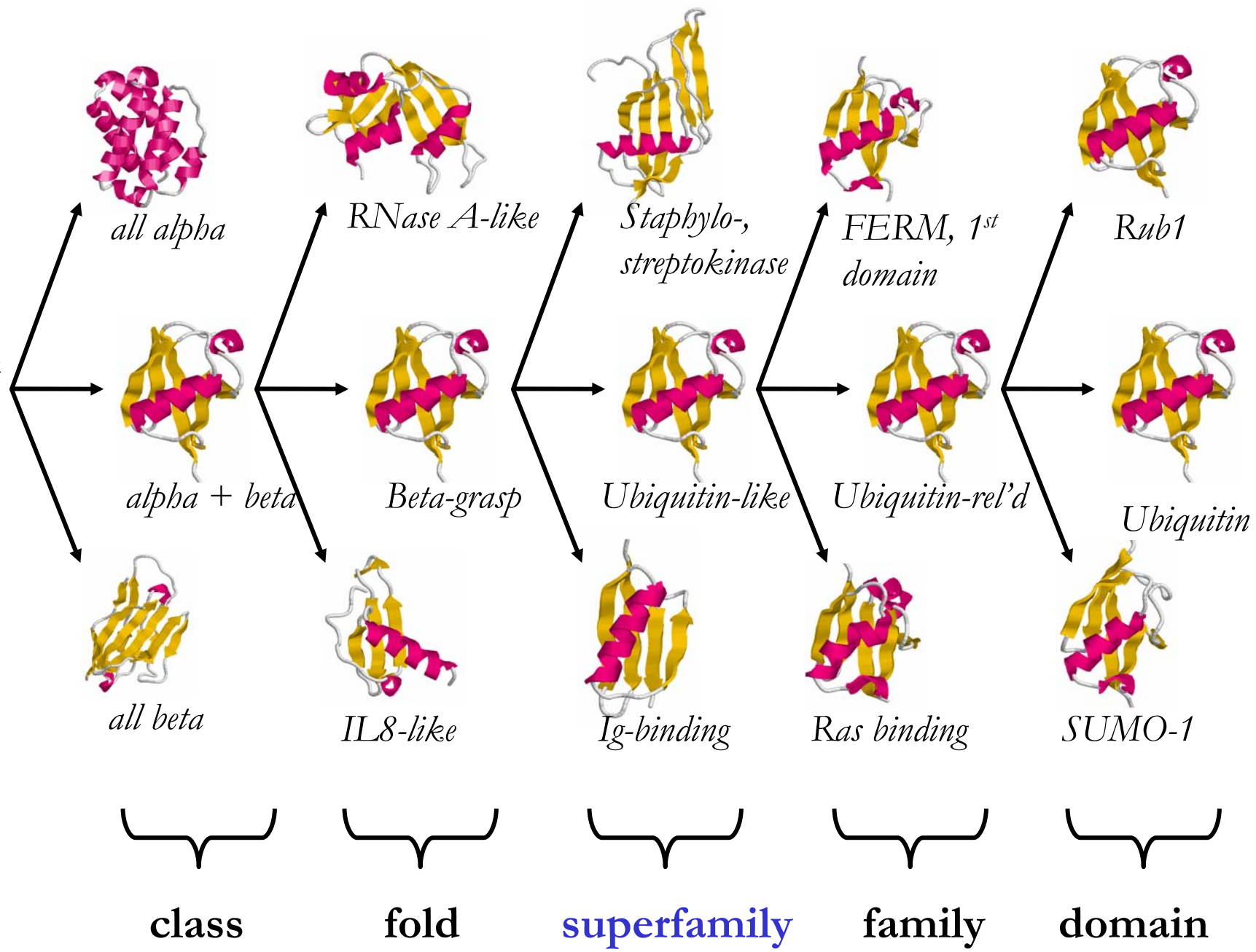
Julian Gough
15.3.07

Gough, J. (2006) Genomic scale sub-family assignment of protein domains. *Nucl. Acids Res.* **34**(13) 3625-3633.

Human body

- Water
- Lipid
- Protein
- trace quantities of other stuff (metals)

Structural
Classific'n
Of
Proteins

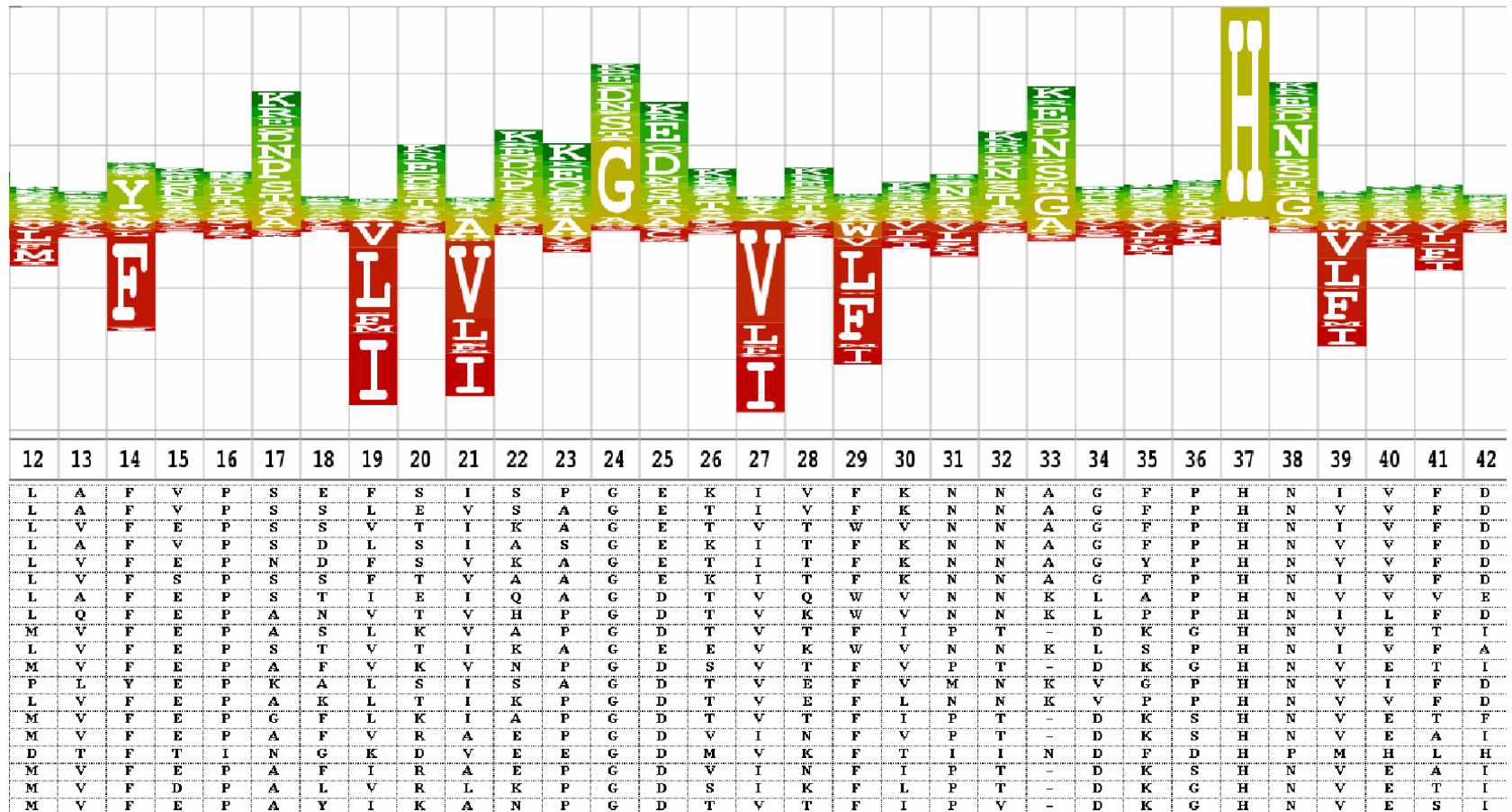


Homology detection

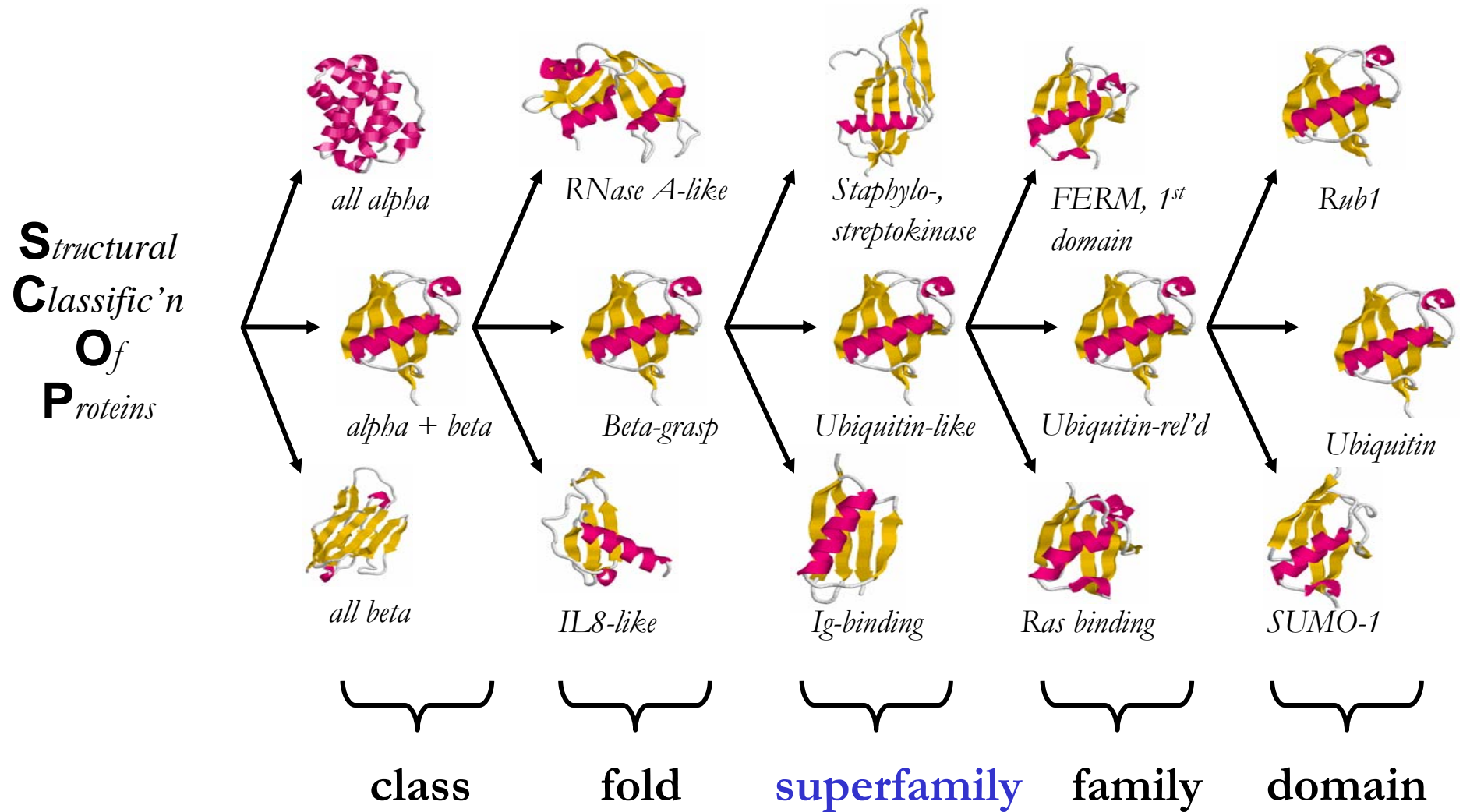
- Mutations cause the amino acid sequence to change during evolution over time
- The structure remains the same
- Different amino acids share similar properties

Profile sequence comparison

- Hidden Markov models



Sub-family classification



Problems

- multi-class classifier needed
- superfamily HMM scores are too general
- HMM models are costly to build
- pairwise methods only find close homologues
- tens of millions of sequences to process
- unseen families

Hybrid method step 1

- Start a sequence for which we've already assigned a superfamily with the HMM
- Extract the alignment of the sequence to the model

```
model:                10      20      30      40      50      60      70      80
sequence  idvllgad---DGSLAFVPSEFSISPGE-----KIVFKNNAG-----FPHNIVFDEDS-IPSGVDASKISMSE

model:                90      100     110     120     130
sequence  EDLLNAKGETFEVAL-SNKGEYSFYC--SPHQGAGMVGKVTVN-----
```

Hybrid method step 2

- For every known member of the superfamily also align it to the HMM

```
model:                10      20      30      40      50      60      70      80
                    |       |       |       |       |       |       |       |
sequence idvllgad---DGSLAFVPSEFSISPGE-----KIVFKNNAG-----FPHNIVFDEDS-IPSGVDASKISMSE
known    .....QIVNSVDTMTLTNANVSPDGFTRAGILVNGVHGPLIRGGKNDNFELNVVNDLDNPTMLRPTSIHWHGLFQRGTNWADGADGVNQ

model:                90      100     110     120     130
                    |       |       |       |       |
sequence EDLLNAKGETFEVAL-SNKGEYSFYC--SPHQGAGMVGKVTVN----
known    CPISPGHAFLYKFTPAGHAGTFWYHSHFGTQYCDGLRGPMVIYDDND
```

Hybrid method step 3

- Calculate the score between the two sequences based on the HMM guided alignment using a substitution matrix and gap penalties

VAL-SNKGEYSFYC--SPHQGAGMVGKVTV
FTPAGHAGTFWYHSHFGTQYCDGLRGPMVI

V	A	L	-	S	N	K	G	E	Y	S	F	Y	C	-	-	S	P	H	Q	G	A	G	M	V	G	K	V	T	V
F	T	P	A	G	H	A	G	T	F	W	Y	H	S	H	F	G	T	Q	Y	C	D	G	L	R	G	P	M	V	I
-	-	-	-	+	-	+	-	+	-	+	+	-	-	-	-	-	-	-	-	+	+	-	+	-	+	-	+	-	+
1	0	3	1	2	1	1	6	1	3	3	3	2	1	1	1	0	1	0	1	3	2	6	2	3	6	1	1	0	3

$$33 - 26 = \mathbf{7}$$

Hybrid method step 4

- After calculating the scores for each member, rank them
- Pick the top score, and the top score to a different family
- Calculate the E-value for the first score using the second score as the null hypothesis

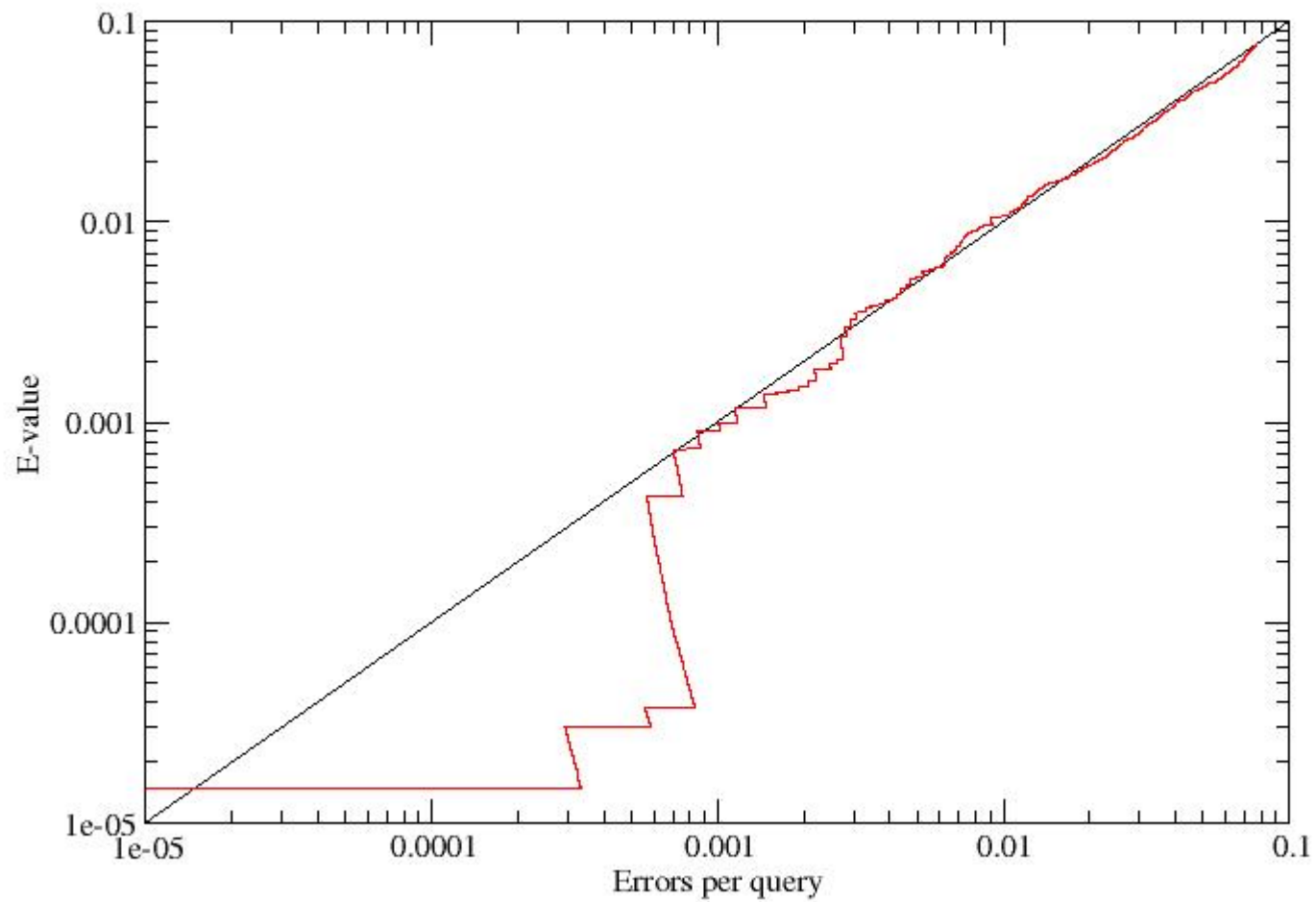
$$E\text{-value} = \frac{K}{1 + e^{(\ln(n_2 e^{-\lambda S_2}) - \ln(n_1 e^{-\lambda S_1}))}}$$

Advantages of the hybrid method

- Works across all homology distances
- Gives a probabilistic score
- Copes with missing families
- Gives a closest homologue
- Computational cost is negligible

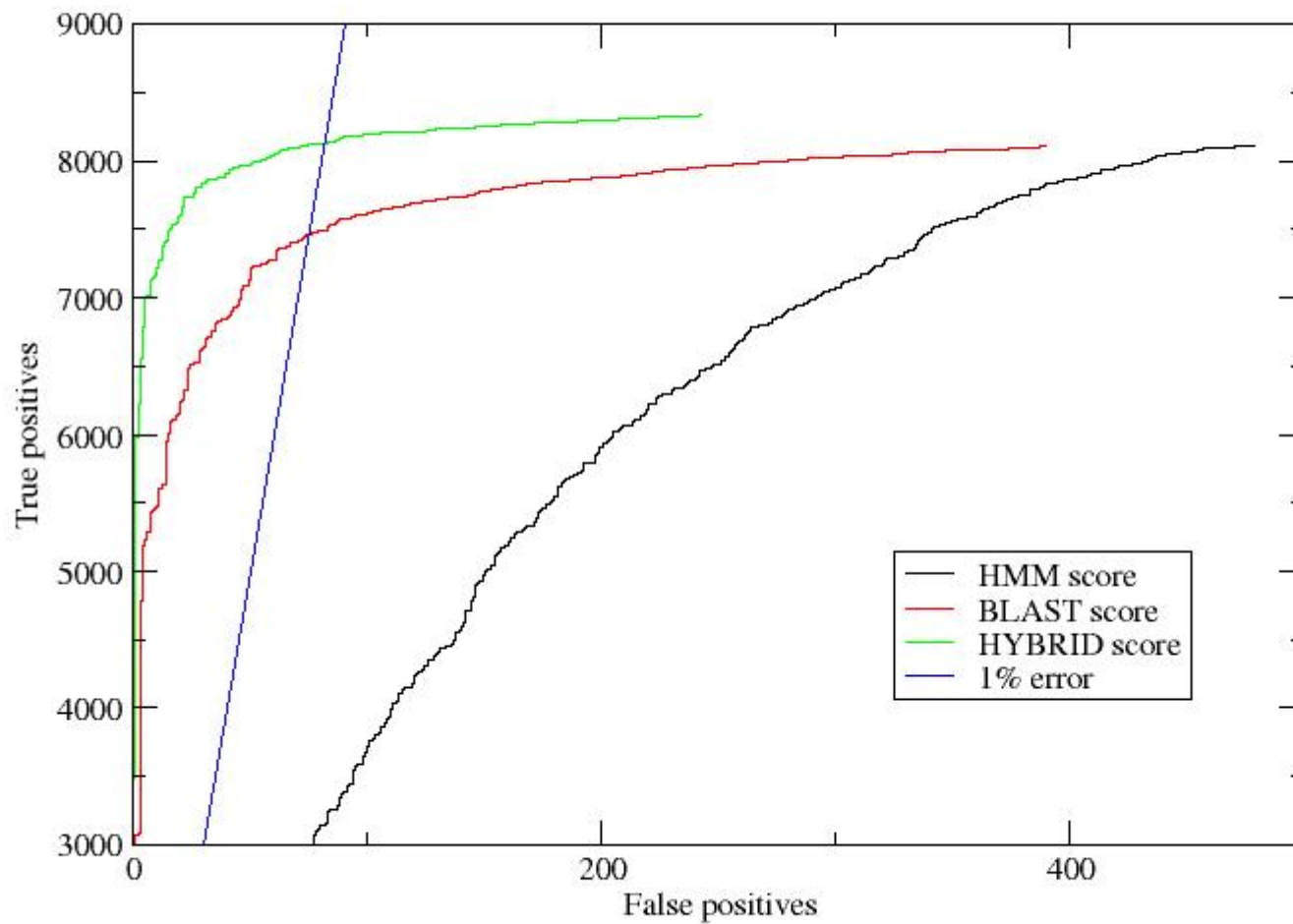
Results

hybrid method E-value calibration against SCOP



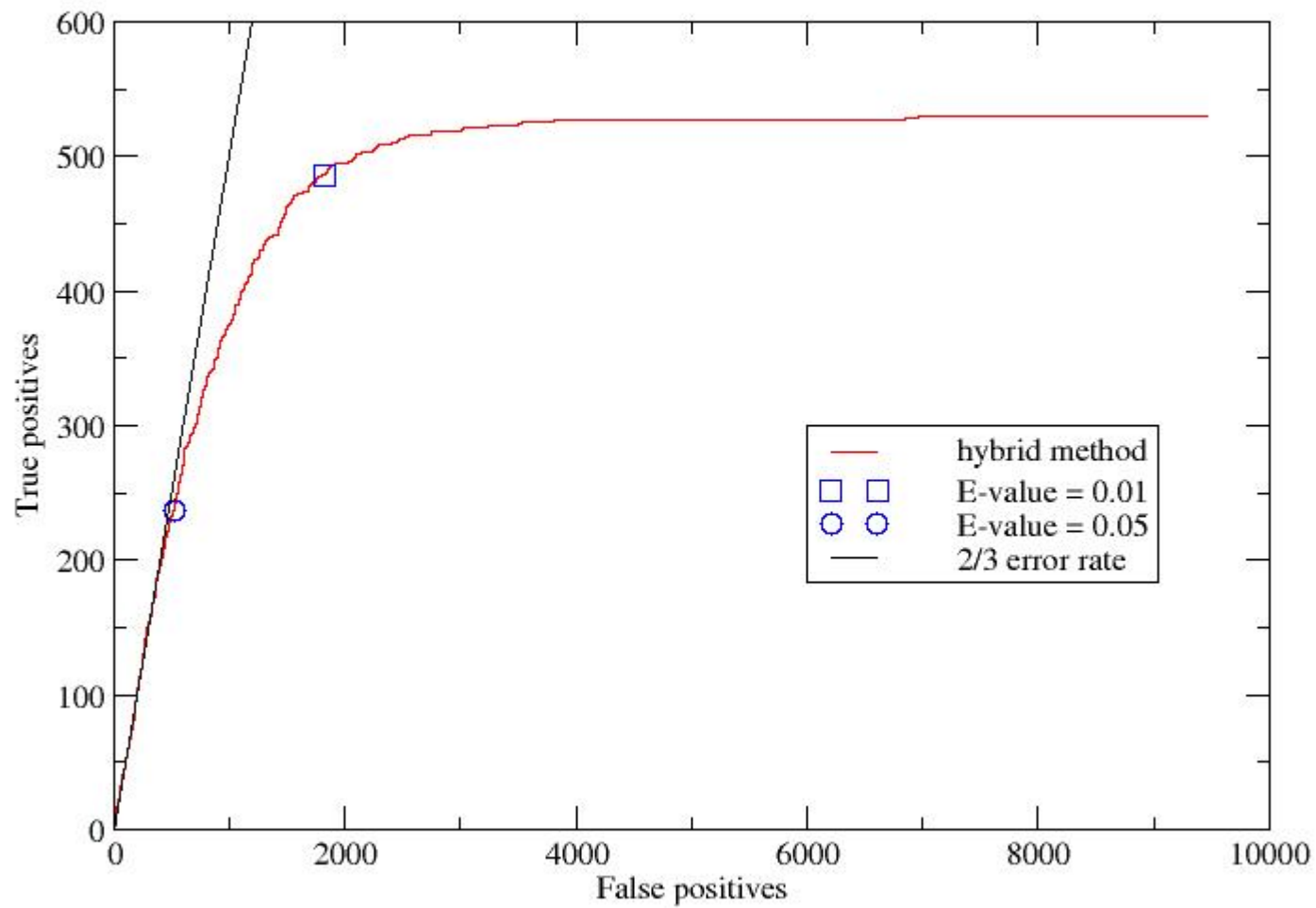
Results

SCOP 1.67 cross-validation of family classification



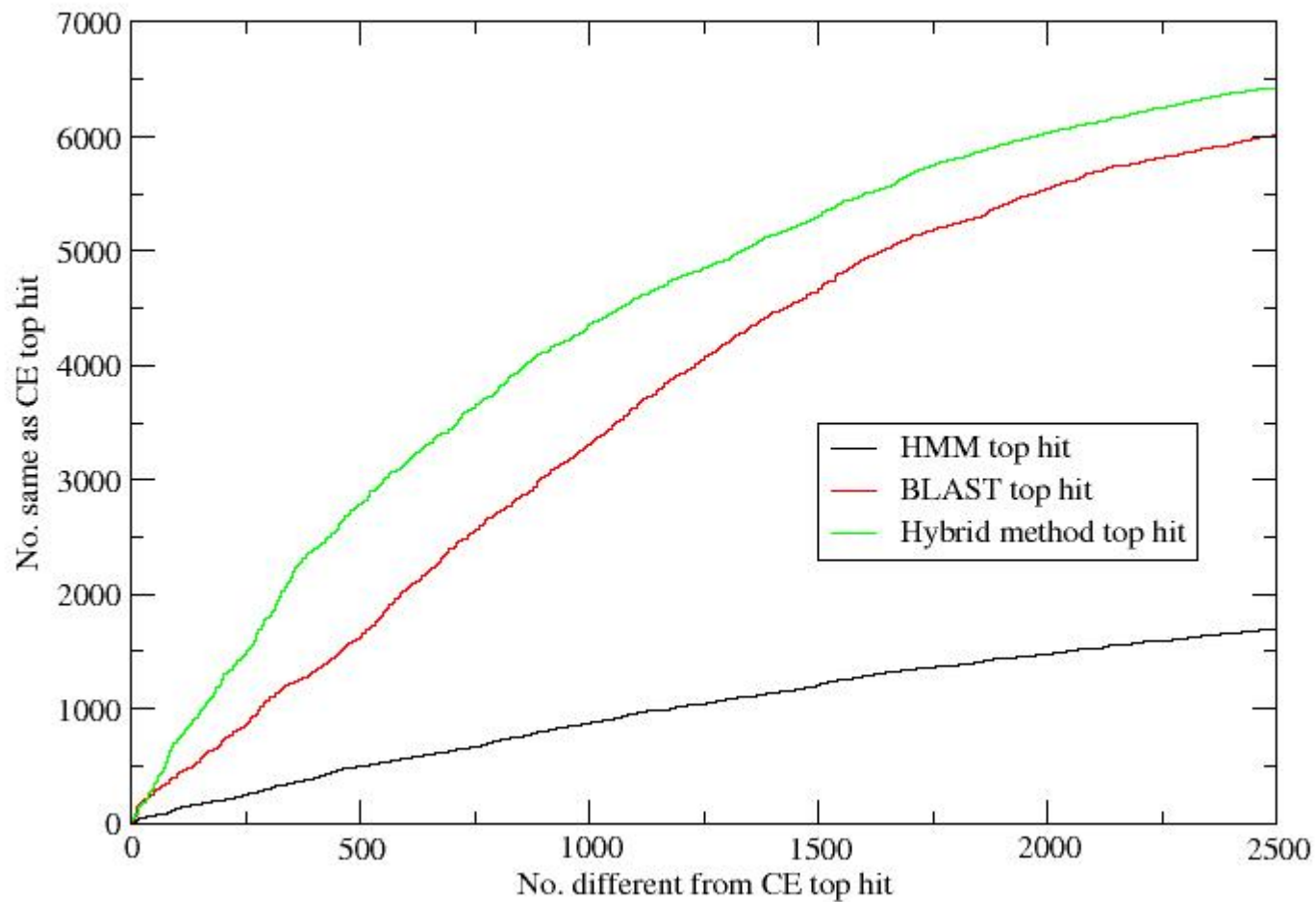
Results

Cross-validation of new family prediction



Results

Agreement of prediction of closest structural homologue with CE



Applications

- Massive annotation
- Protein family evolution