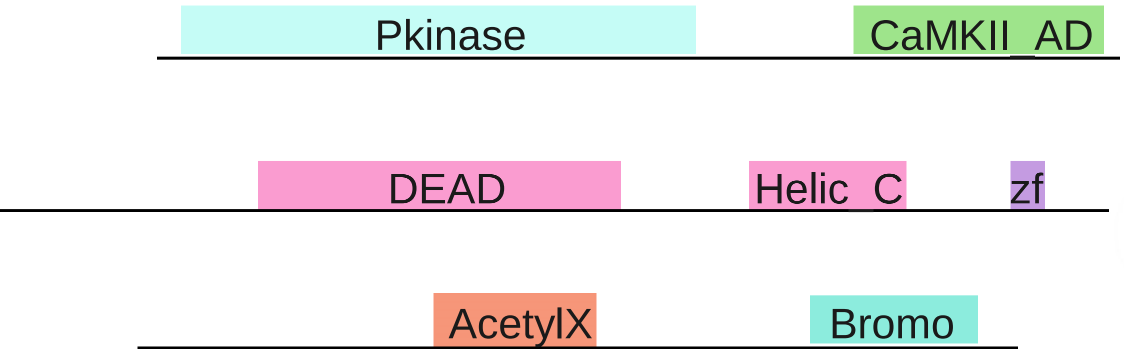
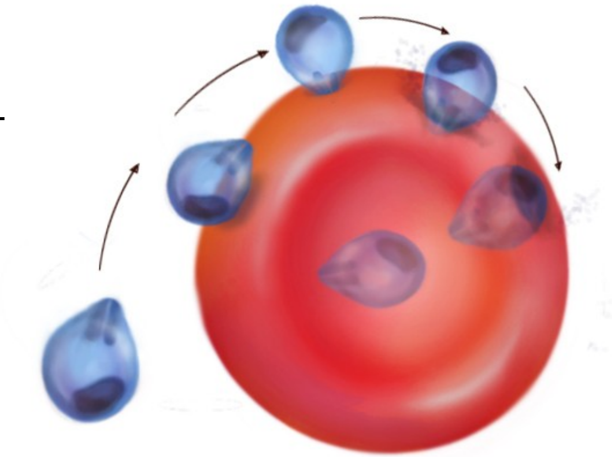
# Forget the *E*-value: family-based *q*-values for protein domain prediction, and empirical error detection



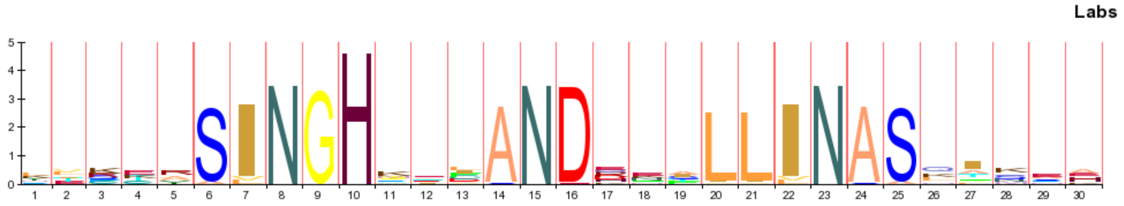


Alejandro Ochoa

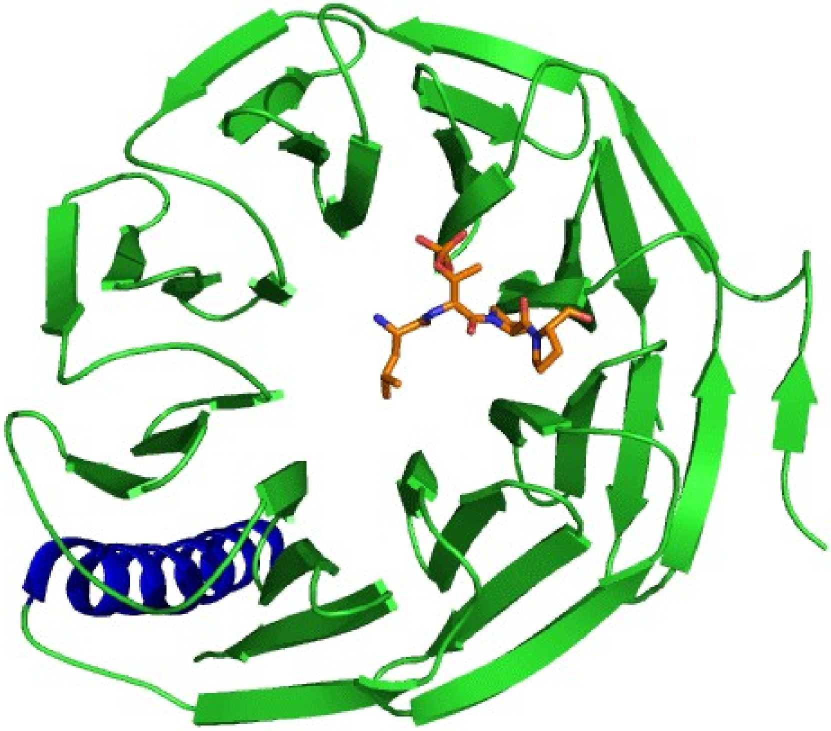
Molecular Biology, Princeton University

NCBI, NIH, 2013-02-25

PRINCETON UNIVERSITY



# Protein domains

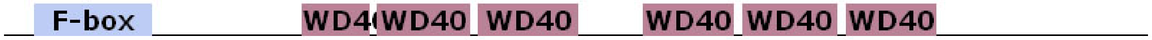


Structure

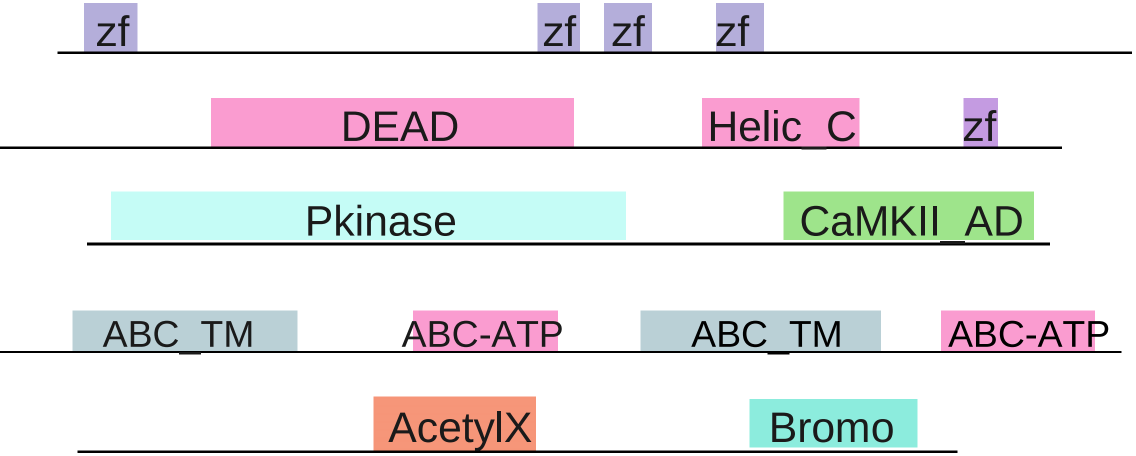
Evolution

Function

Sequence-based domain prediction:



# Why predict domains?



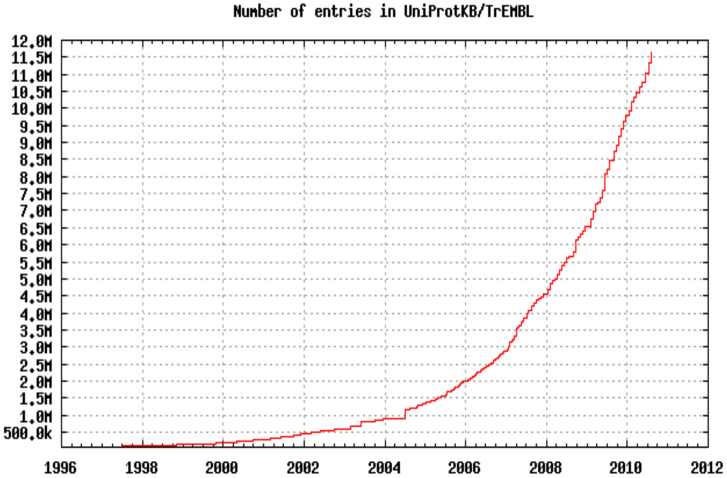
zf zf zf zf

DEAD Helic_C zf

Pkinase CaMKII_AD

ABC_TM ABC-ATP ABC_TM ABC-ATP

AcetylX Bromo



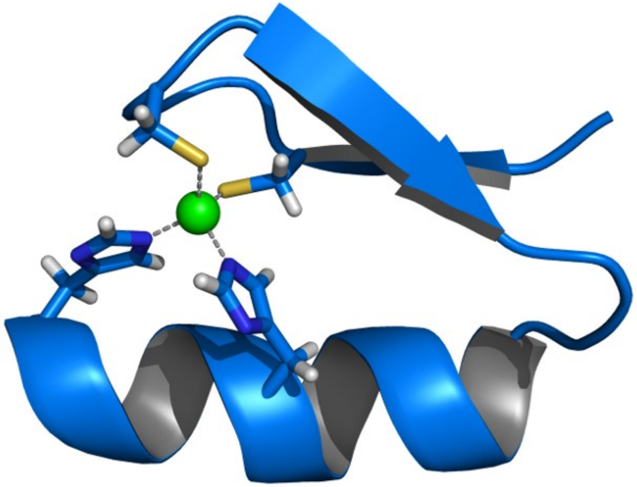
For new sequences, before experiments start…

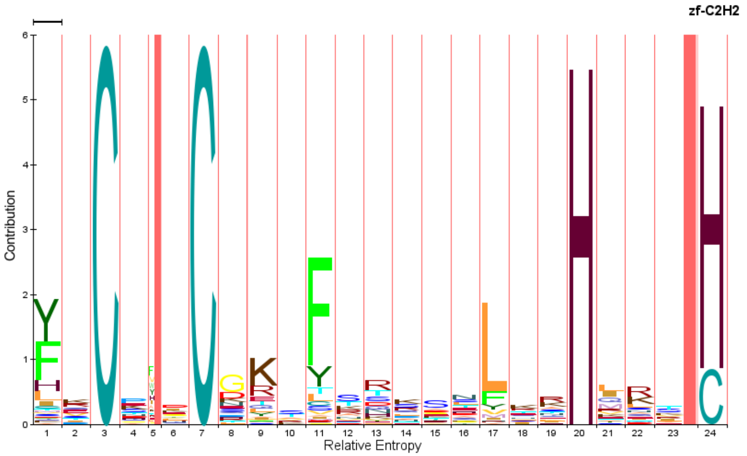
Domains may imply functions

Experimental alternatives are unfeasible as protein databases grow exponentially

# Representing Domains

| | |
|---|---|
| SNAI_DROME/362-385 | YACQ...VCH...KSFSRM...SLLNKHSSS..NC |
| SNAI_XENLA/232-255 | YQCK...SCS...RTFSRM...SLLHKHEET..GC |
| SNAI_MOUSE/236-259 | YQCQ...ACA...RTFSRM...SLLHKHQES..GC |
| ESCA_DROME/426-449 | YSCT...SCS...KTFSRM...SLLTKHSEG..GC |
| SUHW_DROAN/221-243 | HVCG...KCY...KTFRRL...MSLKKHLEF...C |
| TERM_DROME/323-346 | LHCR...RCR...TQFSRR...SKLHIHQKL..RC |
| Z020_XENLA/174-196 | FMCA...DCG...RCFSVS...SSLKYHQRI...C |
| EVI1_HUMAN/217-239 | IKCK...DCG...QMFSTT...SSLNKHRRF...C |
| Z02_XENLA/34-59 | YSCA...DCG...KHFSEK...MYLQFHQKNPSEC |
| EVI1_HUMAN/21-44 | YRCE...DCD...QLFESK...AELADHQKF..PC |
| ZNF10_HUMAN/517-539 | YKCN...QCG...IIFSQN...SPFIVHQIA...H |
| ZNF91_HUMAN/238-260 | YKCE...ECG...KAFKQL...STLTTHKII...C |
| ZFP58_MOUSE/120-142 | IKCE...ECG...KAFSTR...STYYRHQKN...H |
| TRA1_CAEEL/306-331 | YKCEF.ADCE...KAFSNA...SDRAKHQNR..TH |
| ZNF76_HUMAN/345-368 | YTCS...TCG...KTYRQT...STLAMHKRS..AH |
| ZN12_MICSA/106-129 | YRCS...QCG...KAFRRT...SDLSSHRRT..QC |
| LOLA1_DROME/794-817 | YECR...HCG...KKYRWK...STLRRHENV..EC |
| ZNF17_HUMAN/435-457 | YECN...KCG...KFFRYC...FTLNRHQRV...H |
| ZG32_XENLA/34-56 | FVCV...HCG...KGFRDN...YKLSLHLRI...H |
| TF3A_BUFAM/104-128 | YVCYF.ADCG...QQFRKH...NQLKIHQYI...H |
| ZG46_XENLA/146-168 | YVCT...ECG...TSFRVR...PQLRIHLRT...H |
| MZF1_HUMAN/412-434 | FVCG...DCG...QGFVRS...ARLEEHRRV...H |
| ZN239_MOUSE/6-28 | YKCD...KCG...KGFTRS...SSLLVHHSV...H |
| ZSC22_HUMAN/352-374 | YKCG...ECG...KTFSRS...THLTQHQRV...H |
| EGR1_HUMAN/396-418 | FACD...ICG...RKFARS...DERKRHTKI...H |
| SUHW_DROAN/349-373 | YACK...ICG...KDFTRS...YHLKRHQKYS.SC |
| CF2_DROME/485-508 | YTCP...YCD...KRFTQR...SALTVHTTK..LH |
| CF2_DROME/401-423 | YTCS...YCG...KSFTQS...NTLKQHTRI...H |
| KRUP_DROME/306-328 | YTCE...ICD...GKFSDS...NQLKSHMLV...H |
| TYY1_HUMAN/383-407 | YVCPF.DGCN...KKFAQS...TNLKSHILT...H |
| ZG52_XENLA/61-83 | YTCT...QCN...KQFSHS...AQLRAHIST...H |
| TTKB_DROME/538-561 | YPCP...FCF...KEFTRK...DNMTAHVKI..IH |
| ZNF76_HUMAN/285-309 | YTCPE.PHCG...RGFTSA...TNYKNHVRI...H |
| SDC1_CAEEL/145-168 | YMCQ...VCL...TLFGHT...YNLFMHWRT..SC |
| SRYC_DROME/358-380 | YQCD...ICG...QKFVQK...INLTHHARI...H |
| SDC1_CAEEL/270-292 | YFCH...ICG...TVFIEQ...DNLFKHWRL...H |
| TRA1_CAEEL/276-300 | NKCEY.PGCG...KEYSRL...ENLKTHRRT...H |
| ESCA_DROME/370-392 | CKCN...LCG...KAFSRP...WLLQGHIRT...H |
| ... | ... |

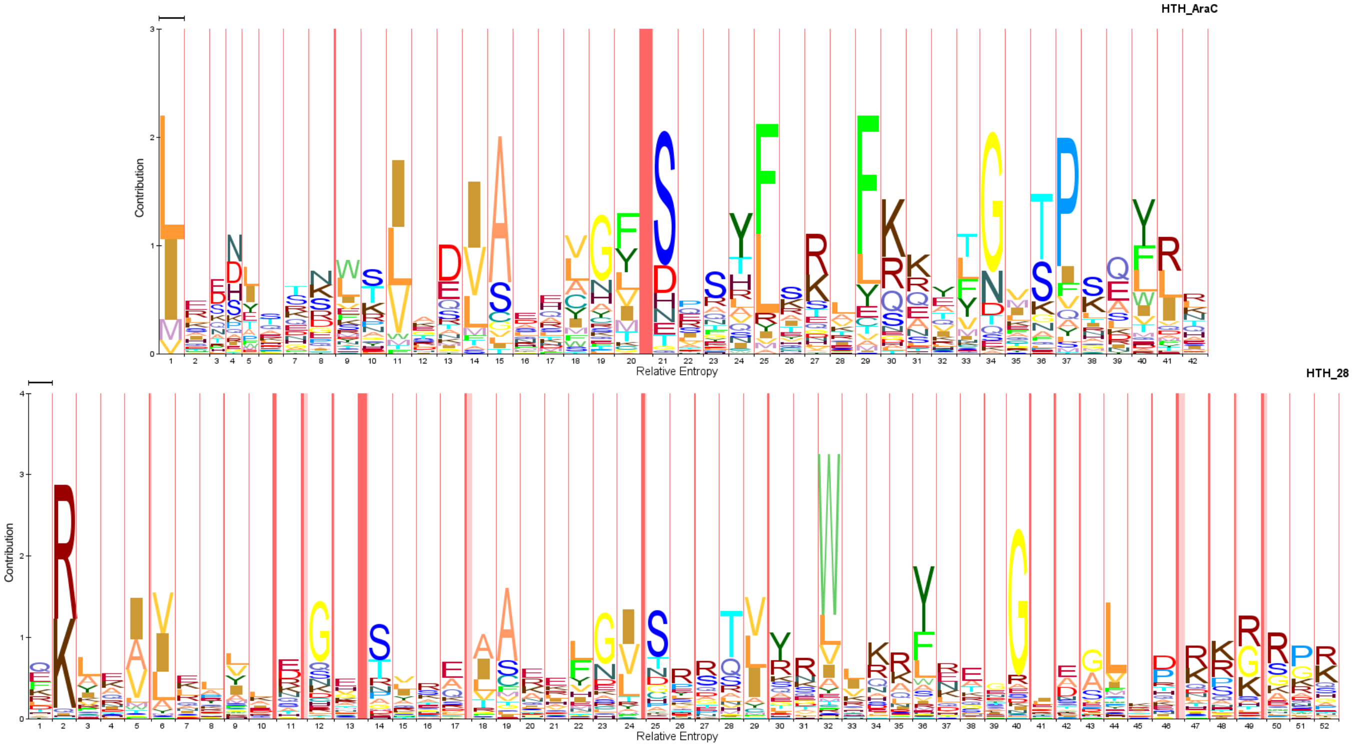




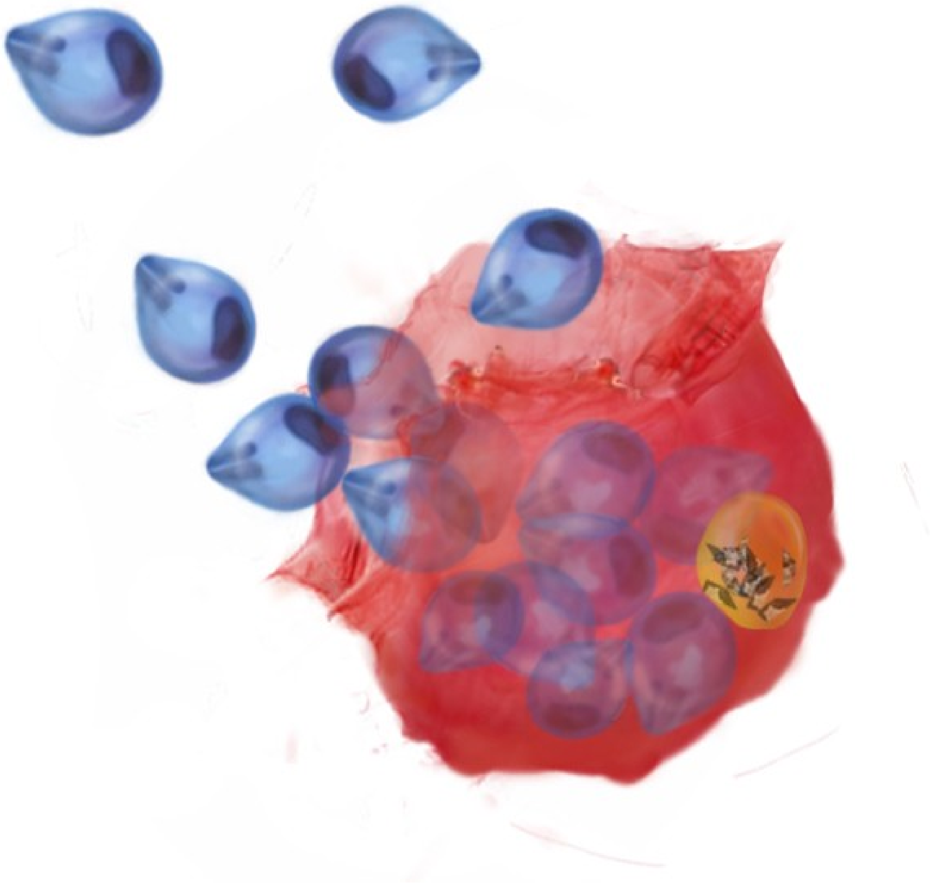
# Databases of Domain Families

This work uses Pfam and HMMER, but theory and results are general

Two members of Pfam “Clan” HTH

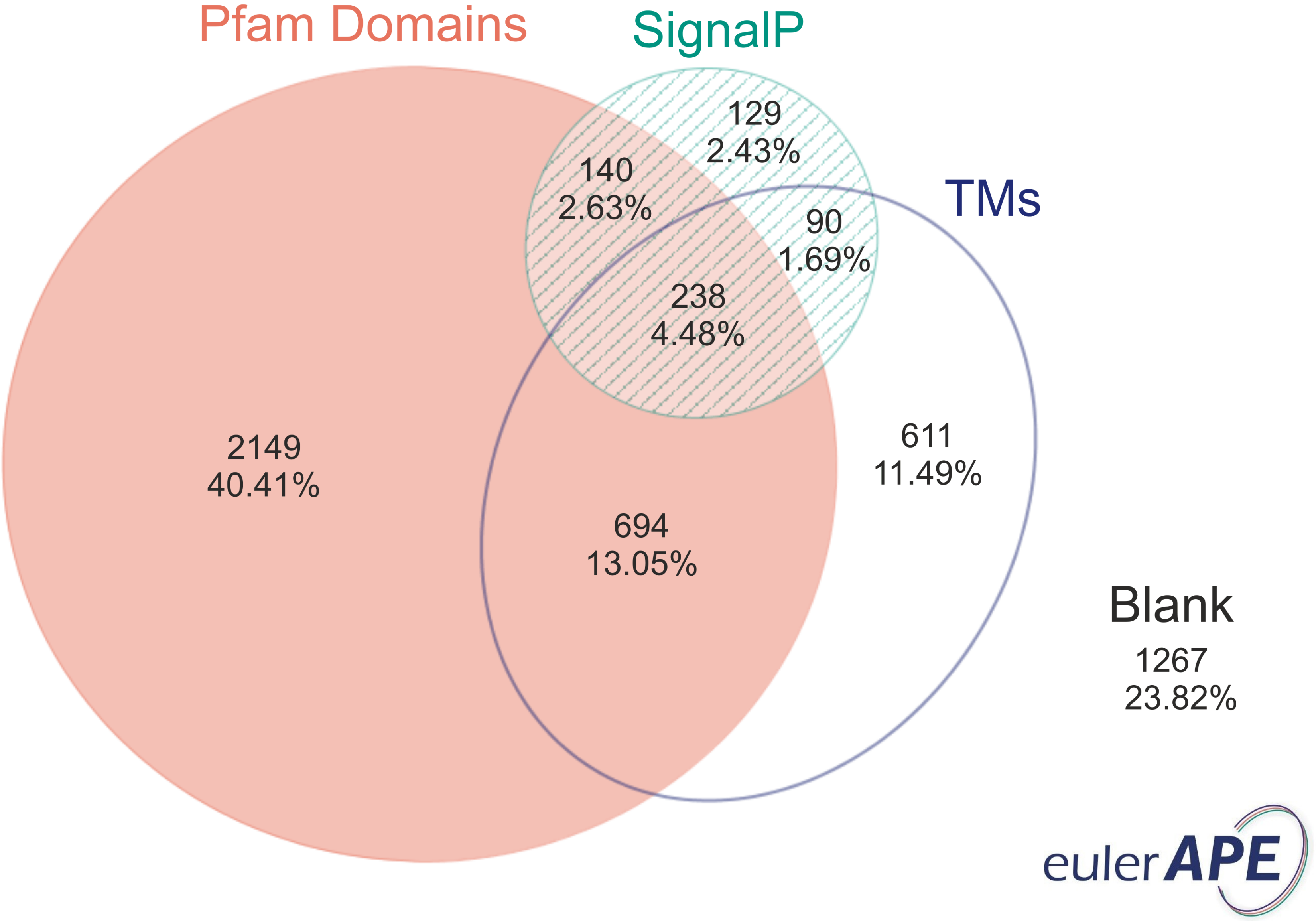


# *Plasmodium falciparum*

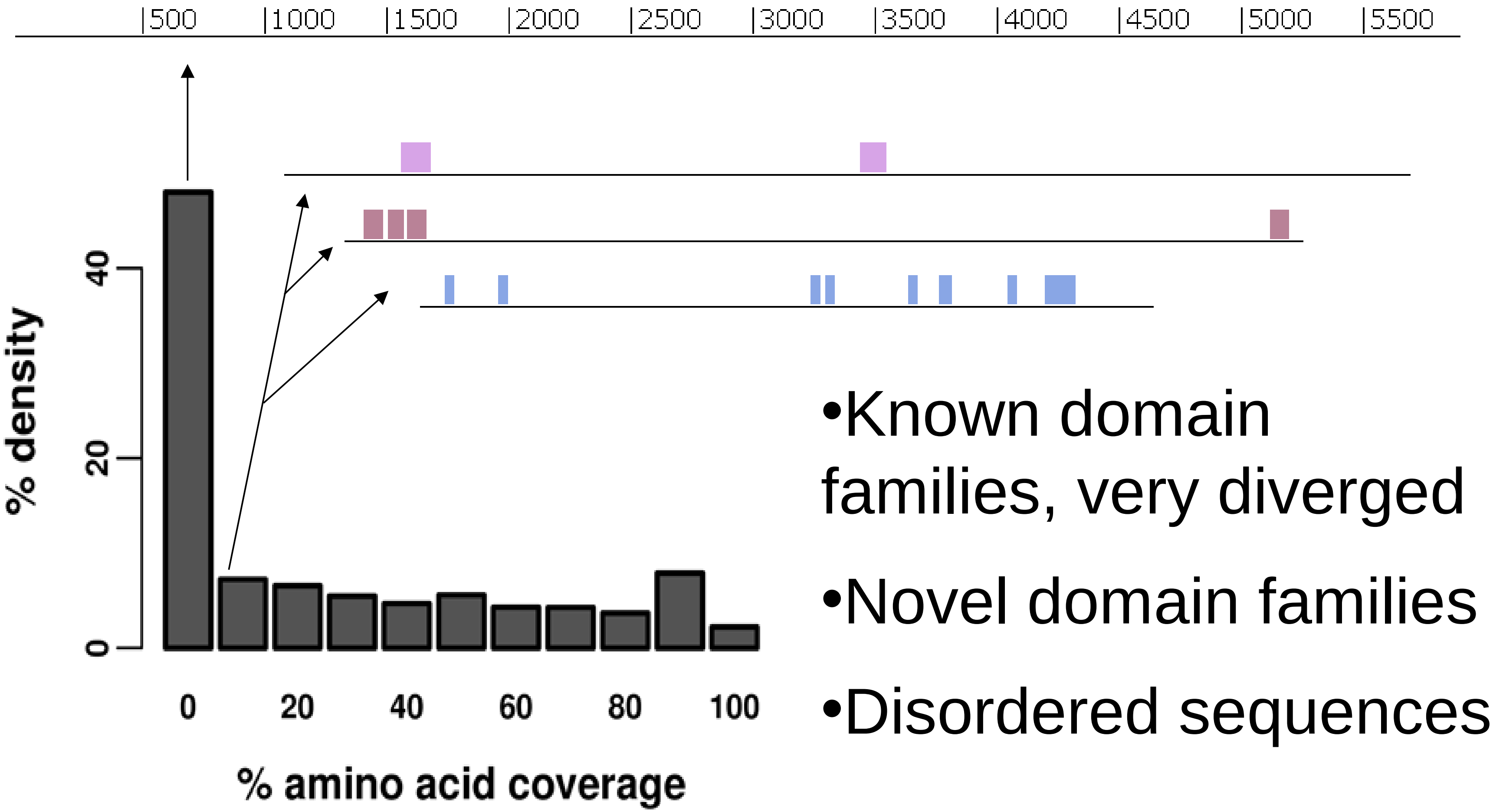


- Malaria
- Information challenges
  - Diverged eukaryote
  - 80% AT-bias
  - Low-complexity regions
- Annotation
  - 5.5K proteins
  - 45% unknown function
    - 20% unknown in yeast
  - 88% of annotations are bioinformatical

# Sequence-based annotation of *Plasmodium falciparum* proteome



# Poor domain coverage of *Plasmodium falciparum*

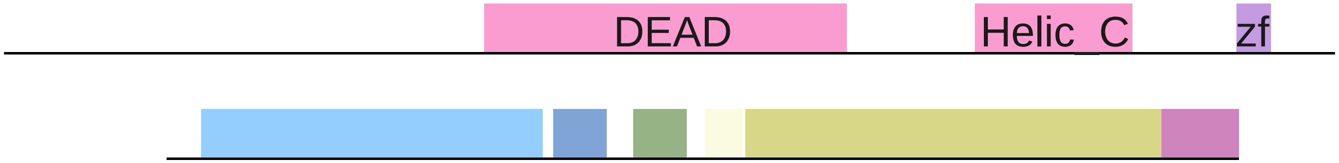


- Known domain families, very diverged
- Novel domain families
- Disordered sequences

# Outline of results

- Domain prediction using context
  - Application to malaria parasite
- Optimal *FDR* control for domains
  - Family-based *q*-values, adapted to domain problem
  - *Local FDR* optimizes problem, but *q*-values are more robust
  - Large problems with coiled coils, transmembrane domains
  - Implications for sequence models

# Domain Prediction Using Context: dPUC



RNA Pol II domains

Background

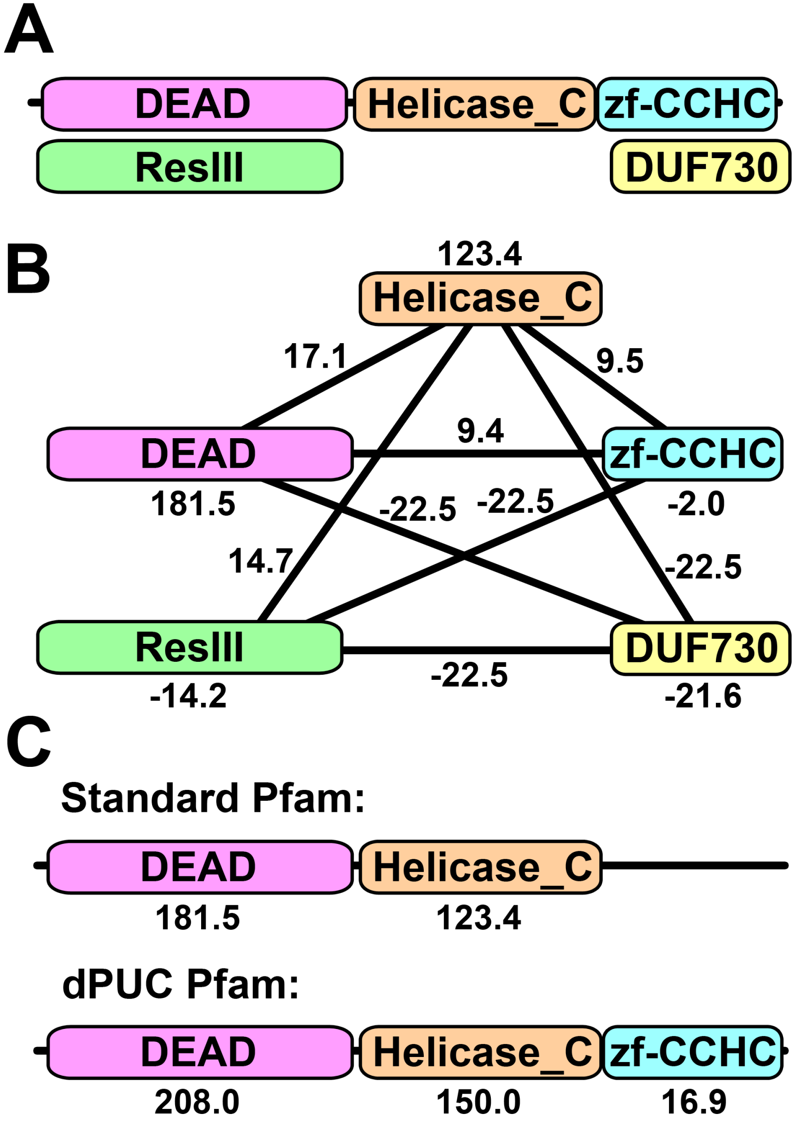
- Domains co-occur in limited combinations
- Domains are scored independently of each other

Idea

- Score domains in combination
- Context + Sequence evidence

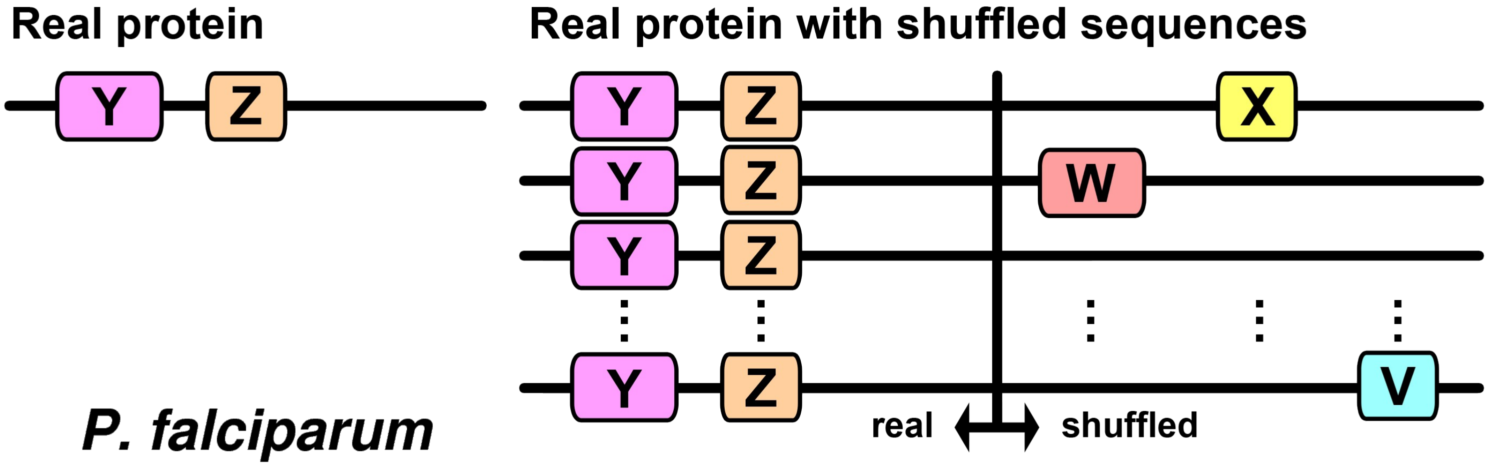
http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# The dPUC method

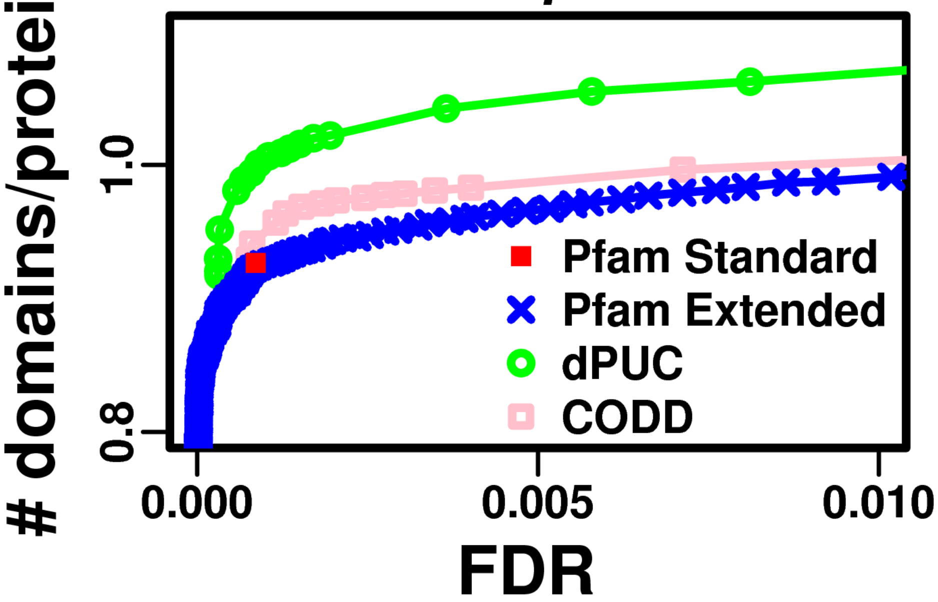


http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# Improved signal to noise

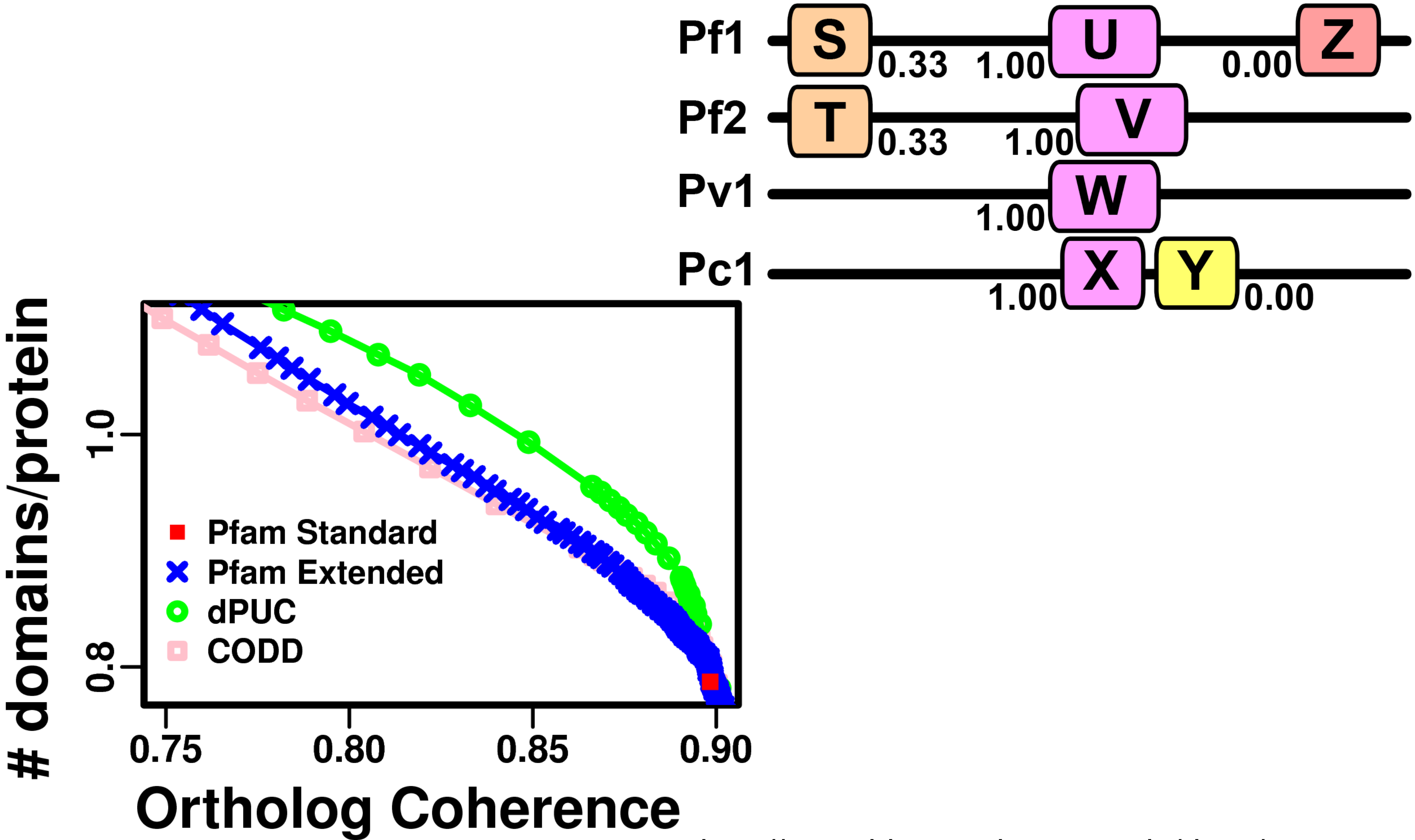


*P. falciparum*



http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# Improved ortholog coherence on *Plasmodium* species

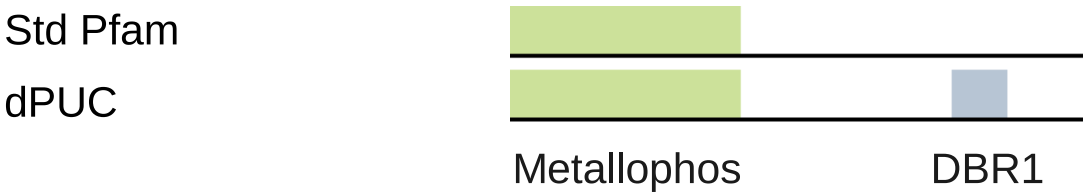


http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# New predictions

Phosphatase -> RNA lariat debranching enzyme

*P. falciparum*



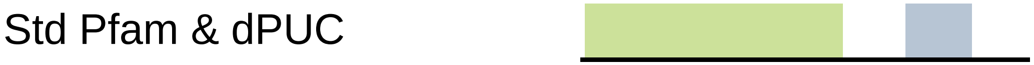
Std Pfam

dPUC

Metallophos

DBR1

*S. cerevisiae*

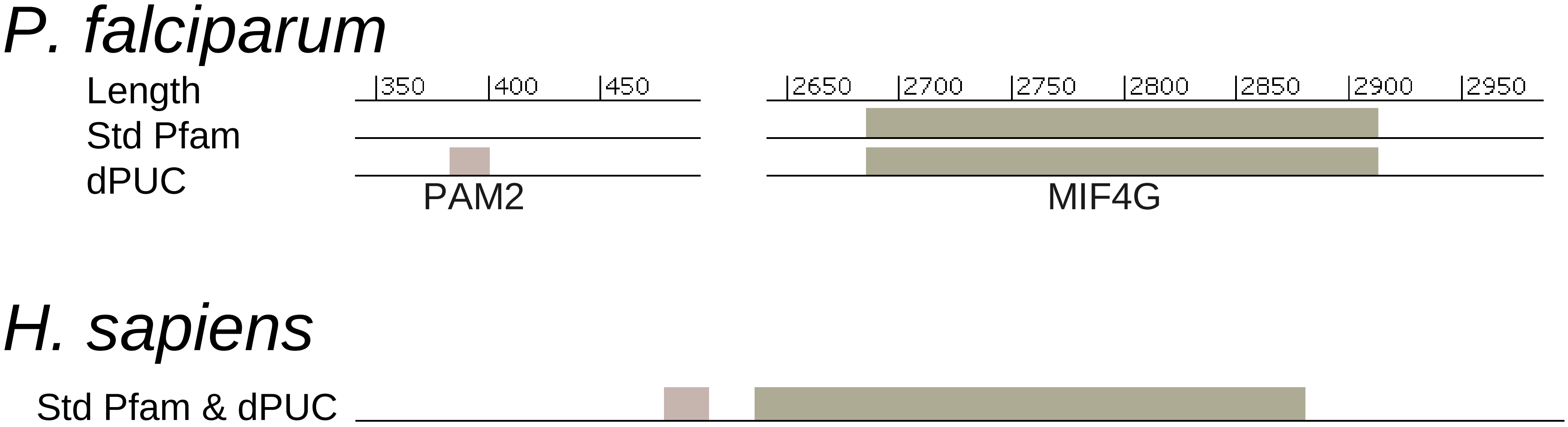


Std Pfam & dPUC

http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# New predictions

MIF4G domain-containing protein ->
Poly-A binding protein-interacting protein 1

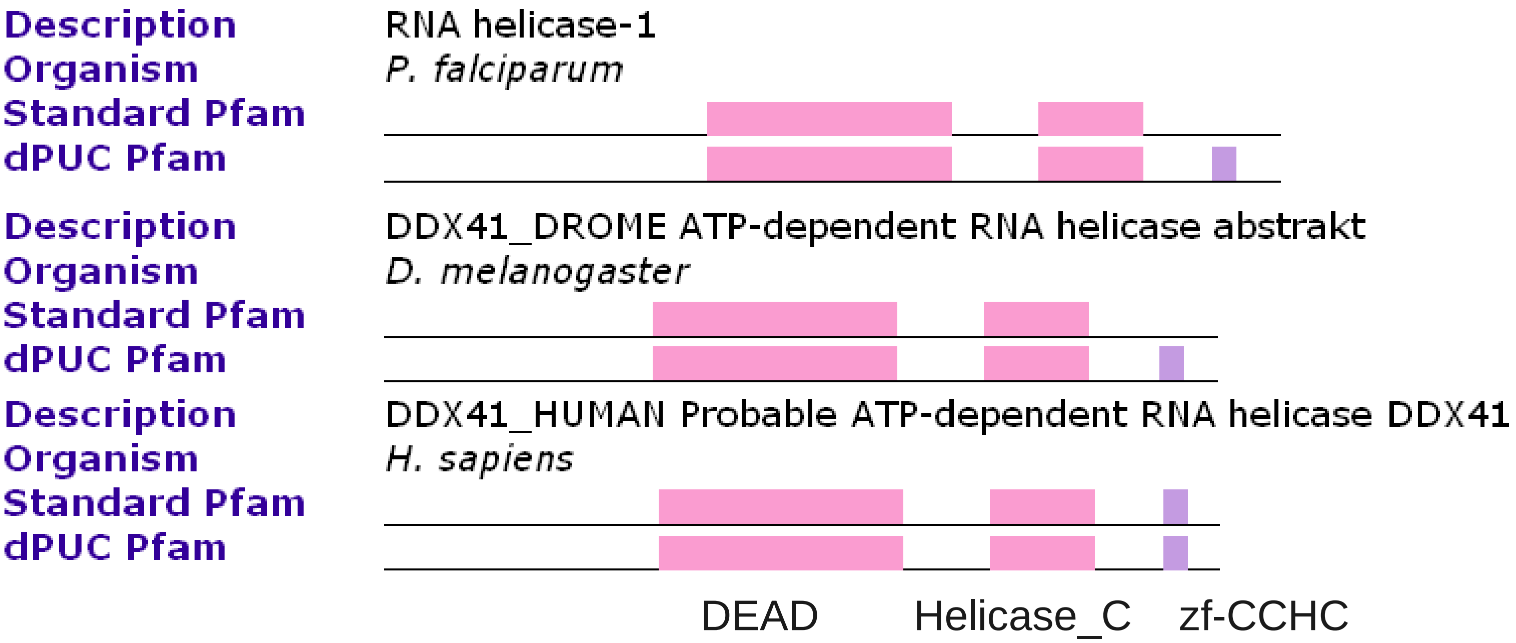


Interested malaria curator:
Hagai Ginsburg,
Hebrew U of Jerusalem

http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# New predictions

RNA helicase -> mRNA sequestration



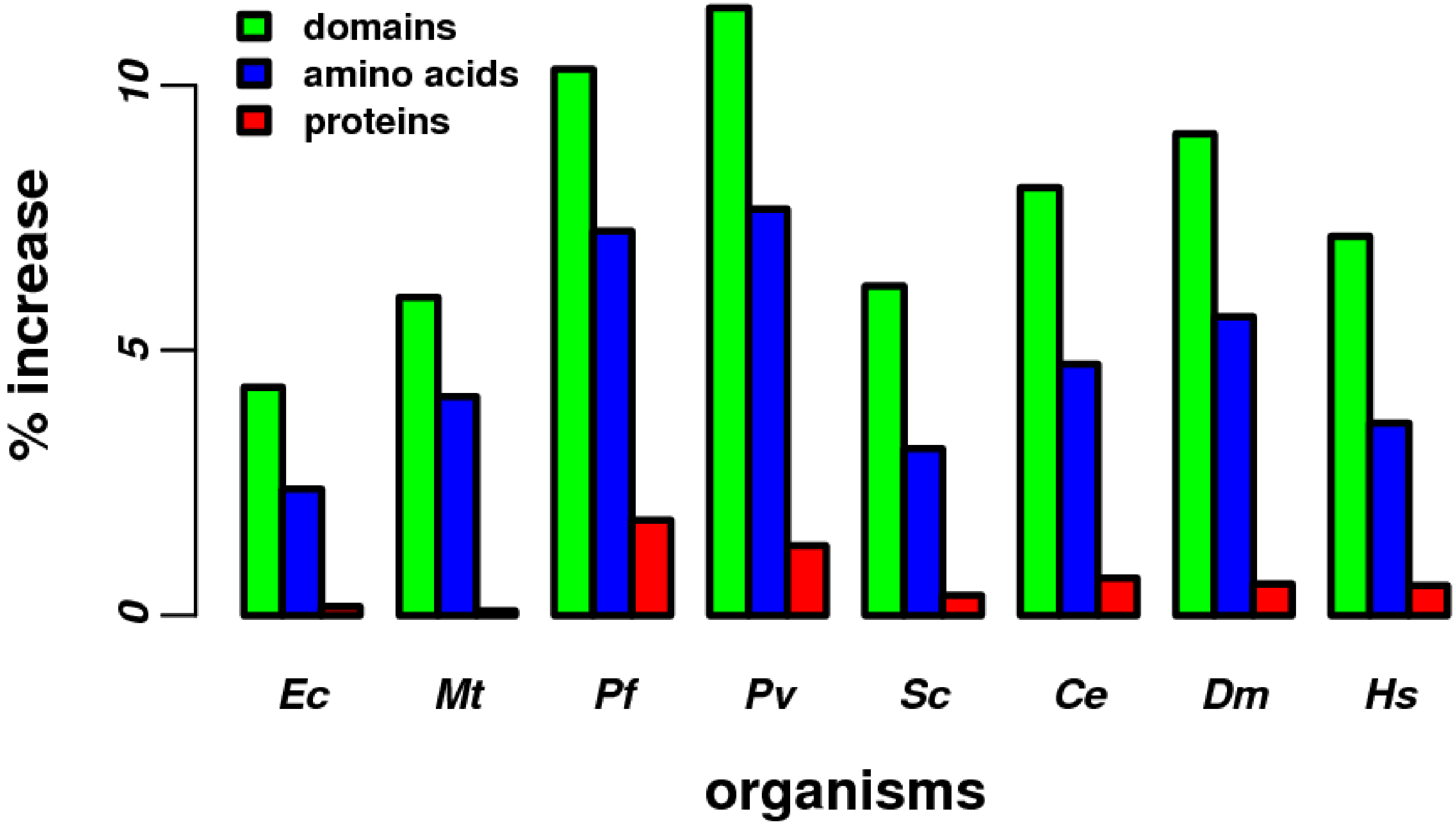
**Description** RNA helicase-1
**Organism** *P. falciparum*
**Standard Pfam**
**dPUC Pfam**

**Description** DDX41_DROME ATP-dependent RNA helicase abstrakt
**Organism** *D. melanogaster*
**Standard Pfam**
**dPUC Pfam**

**Description** DDX41_HUMAN Probable ATP-dependent RNA helicase DDX41
**Organism** *H. sapiens*
**Standard Pfam**
**dPUC Pfam**

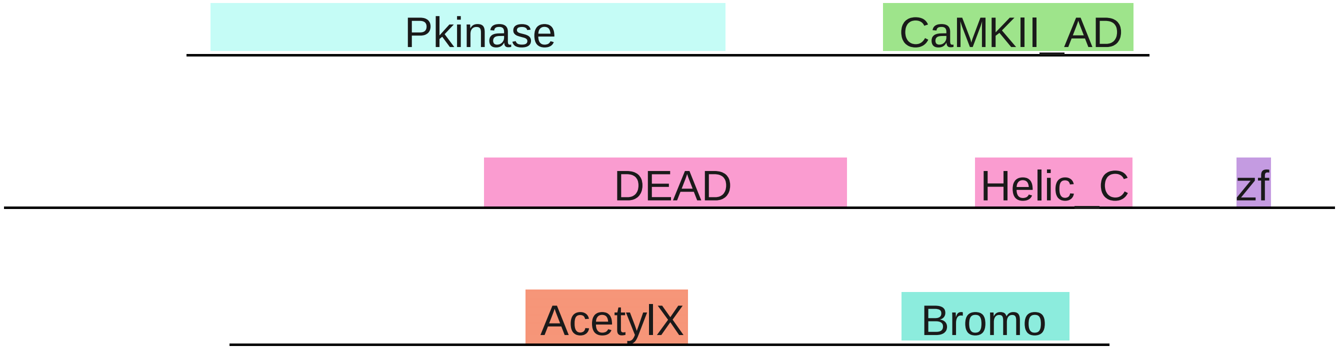
http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# dPUC increases coverage



http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# Domain context



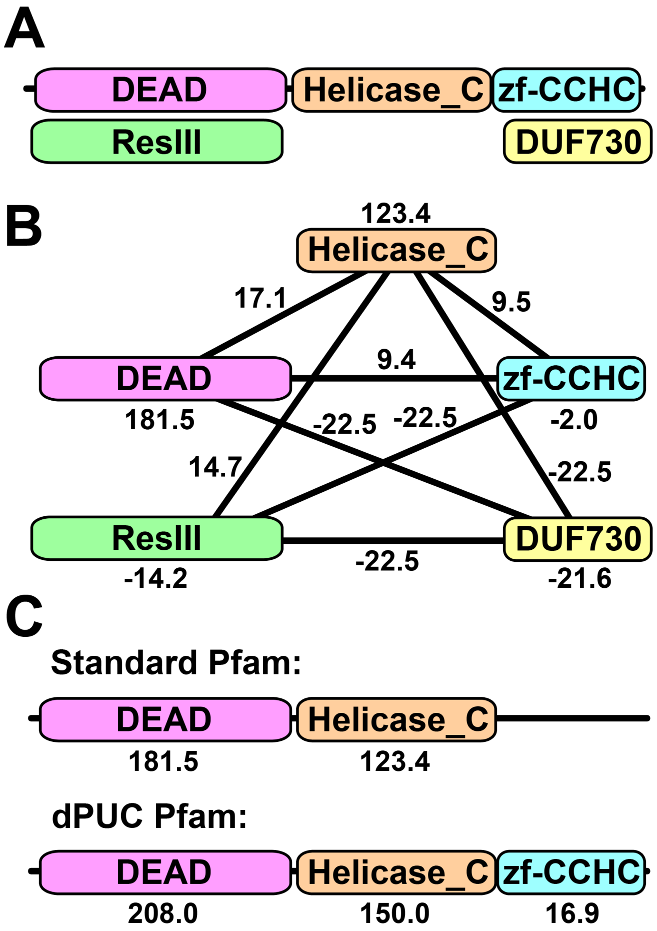
Complements sequence evidence

Improves domain predictions

Works best on diverged organisms

http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# The dPUC method: problems



Domain scores are normalized by curated thresholds

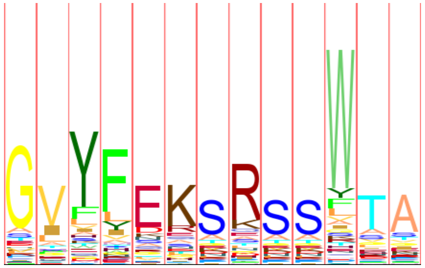
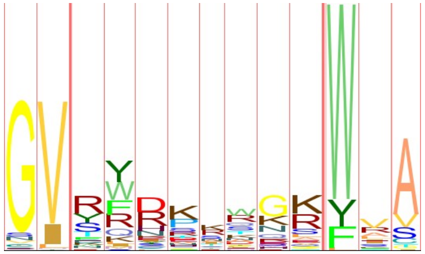
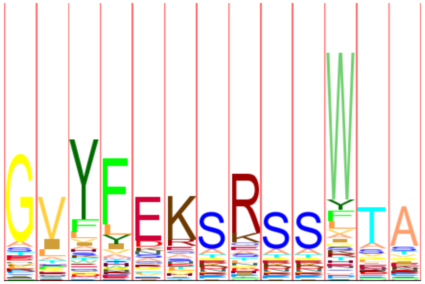
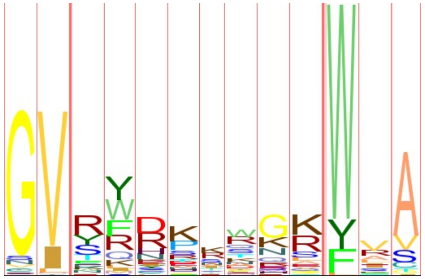
Why do *E*-values perform worse?

http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

# Outline of results

- Domain prediction using context
  - Application to malaria parasite
- Optimal *FDR* control for domains
  - Family-based *q*-values, adapted to domain problem
  - *Local FDR* optimizes problem, but *q*-values are more robust
  - Large problems with coiled coils, transmembrane domains
  - Implications for sequence models

# The *E*-value: the statistic of sequence analysis

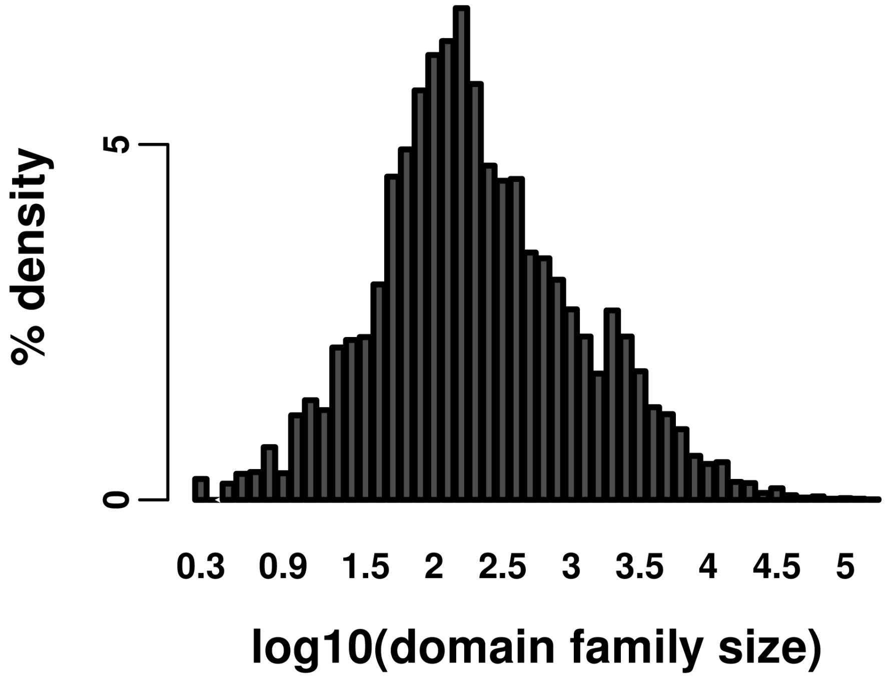
| Query Type | Database Type | Software |
|---|---|---|
| GVYFEKSRSSWTA | GVRQRKNSNKWVS | **FASTA (*Z* 1985)**<br>**BLAST (*E* 1990)** |
|  | GVRQRKNSNKWVS | **PSI-BLAST (*E* 1997)** |
| GVYFEKSRSSWTA |  | **SAM (*Z* 1994, ~*E* 2005)**<br>**HMMer (bit 1995, ~*E* 1998, *E* 2008)**<br>**Pfam DB (curated 1997)**<br>**IMPALA, RPS-BLAST (*E* 1999)** |
|  |  | **PRC (bit 2002, ~*E* 2004)**<br>**HHsearch (~*E*,*P* 2004)**<br>**CORAL (*E* 2009)** |

**First genomes**
**1976: virus**
**1995: bacteria**
**1996: yeast**
**1998: worm**
**2000: human**

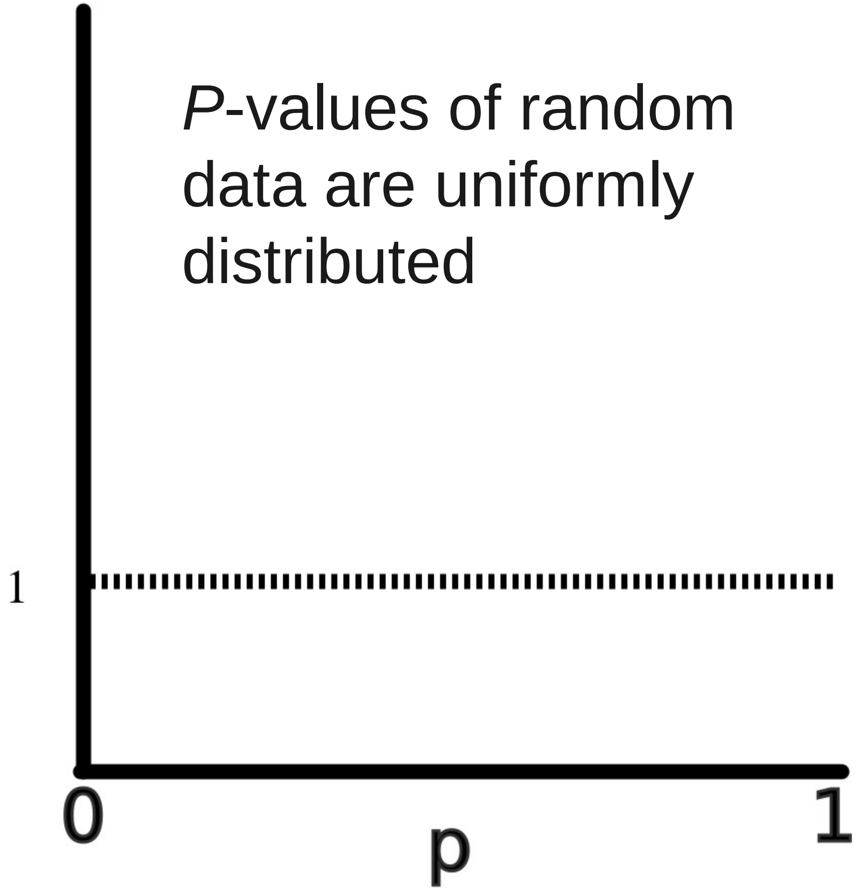
***FDR*: BH 1995, *q* 2002**

# *FDR* = *E*/*n*

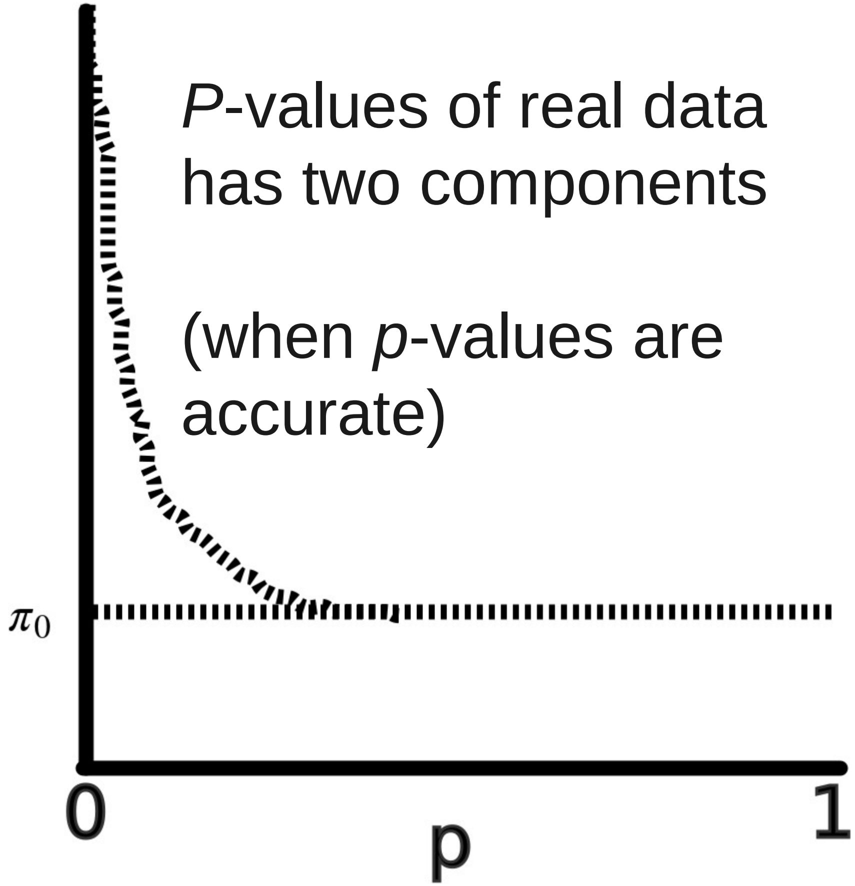
*FDR* = average posterior error probability



# Computing *q*-values



*P*-values of random data are uniformly distributed

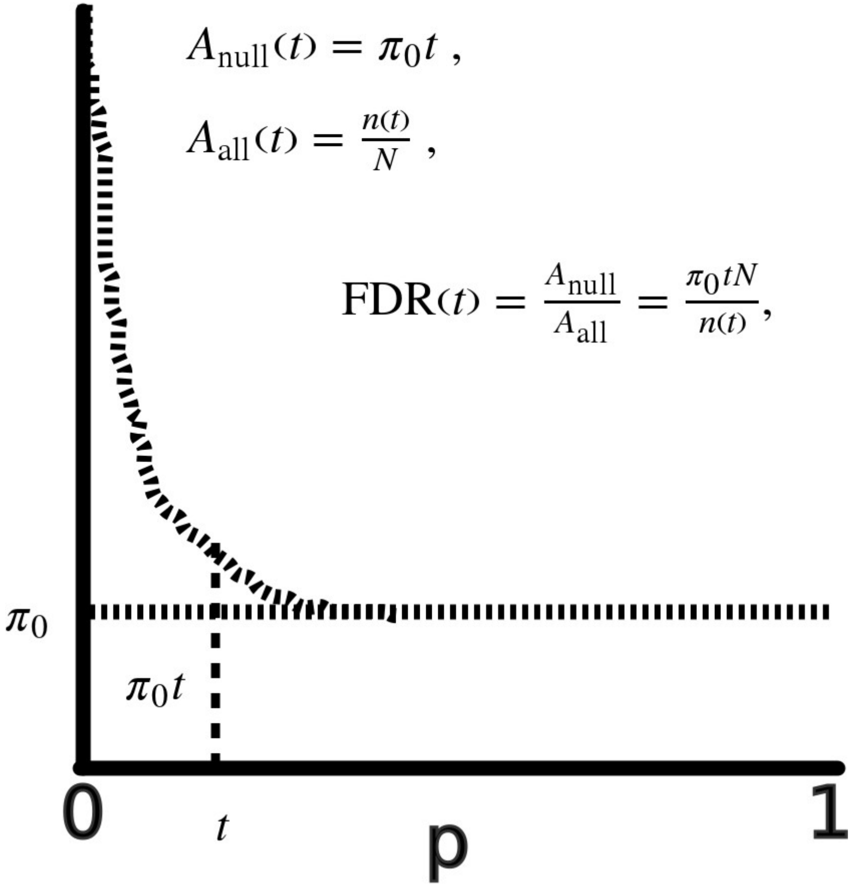


*P*-values of real data has two components

(when *p*-values are accurate)

Step 1: estimate $\pi_0$ (proportion of data that is false)

# Computing *q*-values



$$A_{\text{null}}(t) = \pi_0 t\ ,$$

$$A_{\text{all}}(t) = \frac{n(t)}{N}\ ,$$

$$\text{FDR}(t) = \frac{A_{\text{null}}}{A_{\text{all}}} = \frac{\pi_0 t N}{n(t)},$$

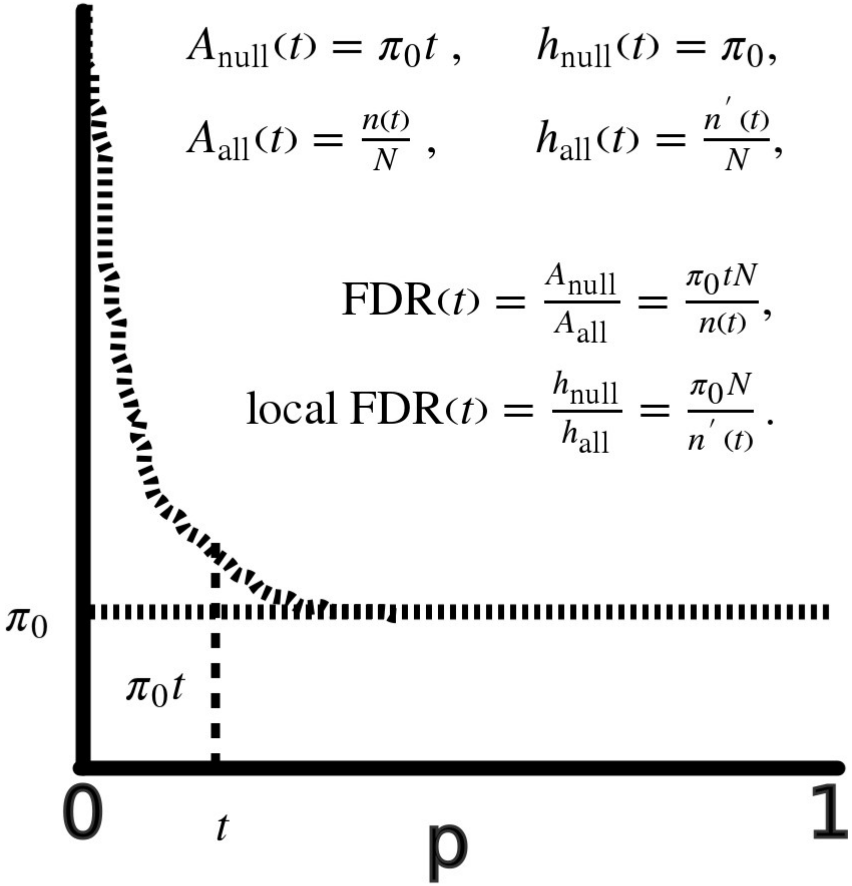
Step 2: Directly estimate *FDR*(*t*) for all thresholds *t*

- – *N* = # tests
- – *n(t)* = # sig tests

Step 3: Ensure monotonicity

$q(p) = \min_{t;p \leq t} FDR(t)$

# *FDR* and *local FDR*



$$A_{\text{null}}(t) = \pi_0 t \,, \qquad h_{\text{null}}(t) = \pi_0,$$

$$A_{\text{all}}(t) = \frac{n(t)}{N} \,, \qquad h_{\text{all}}(t) = \frac{n'(t)}{N},$$

$$\text{FDR}(t) = \frac{A_{\text{null}}}{A_{\text{all}}} = \frac{\pi_0 t N}{n(t)},$$

$$\text{local FDR}(t) = \frac{h_{\text{null}}}{h_{\text{all}}} = \frac{\pi_0 N}{n'(t)}.$$

*Local FDR* =
Posterior Error
Probability (*PEP*)

*FDR* = average *PEP*
of significant
predictions

# *Local FDR*s optimize domain prediction

Find domain family thresholds $t_i$ (for each family *i*) to maximize predictions

$$M = \sum n_i(t_i) \, ,$$

while constraining the combined *FDR* of all families to *Q*

$$Q \geq \frac{\sum \pi_{0,i} \cdot t_i \cdot N_i}{\sum n_i(t_i)} = \frac{\sum FDR_i(t_i) \cdot n_i(t_i)}{\sum n_i(t_i)} \, .$$

Equal family *local FDR*s solve this optimization!

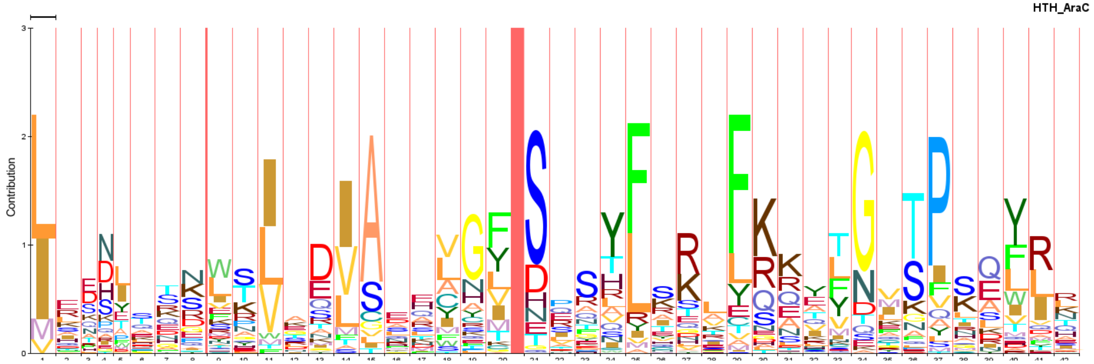
# *Q*-values for domains

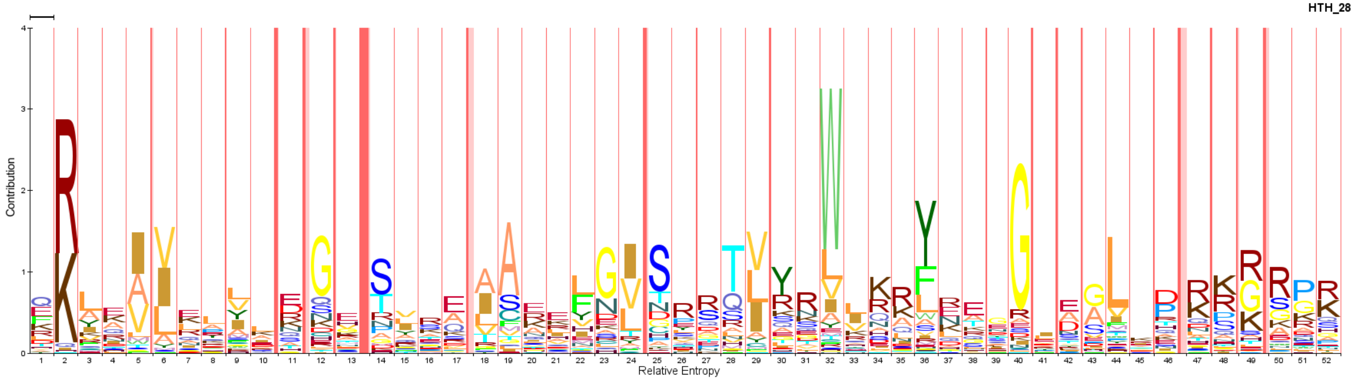
Introduced domain-specific features

- Incomplete list of *p*-values (HMMER3, BLAST and descendants)
- Correction for domain overlaps (heuristic)

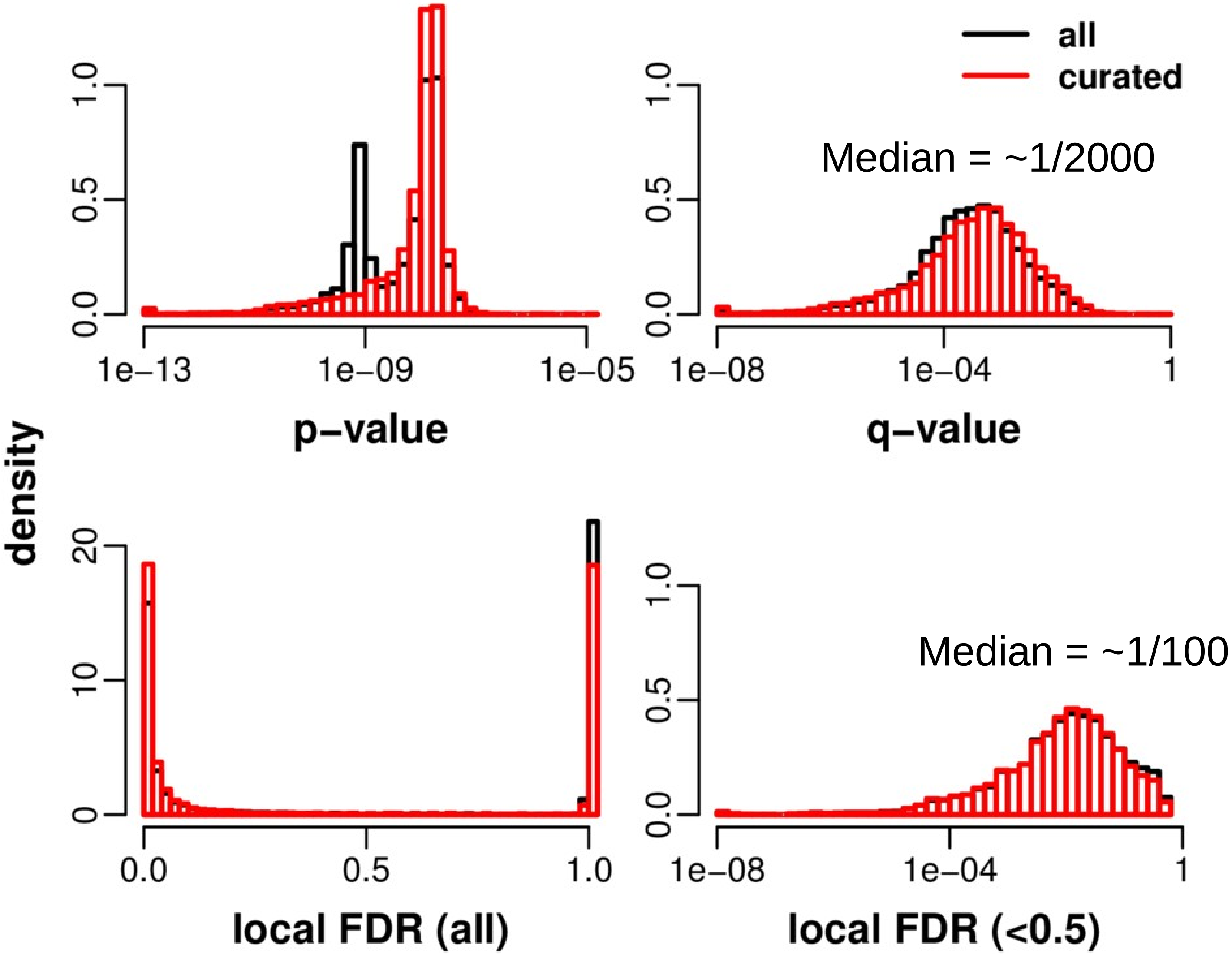
Two members of Pfam “Clan” HTH

$$FDR = E/n$$





# Pfam curated thresholds



Median = ~1/2000

Median = ~1/100

# Empirical null models

***FDR = #FP / ( #FP + #TP )***

## 2nd order Markov Random Sequences

- **FP if domain came from random sequence**

Improved from Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

## Ortholog Set Coherence

- **FP if orthologs don't predict any homologous domains**

Improved from Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

## Clan Overlap

- **FP if domain overlaps stronger non-homologous domain**

Inspired/adapted from S. Eddy (p.c., 2012).

## Context Coherence

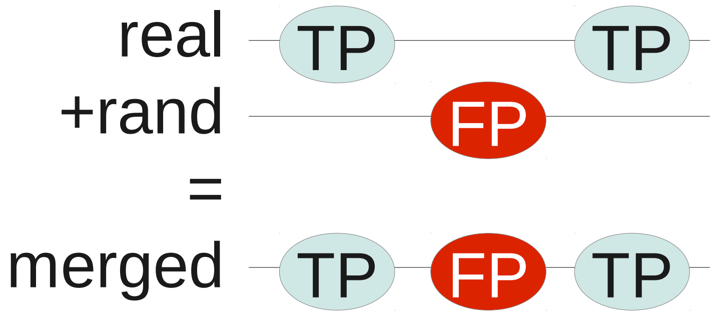
- **FP if domain doesn't co-occur with any stronger domains**

Inspired/adapted from Terrapon, *et al*. BMC Bioinformatics 13, 67 (2012).

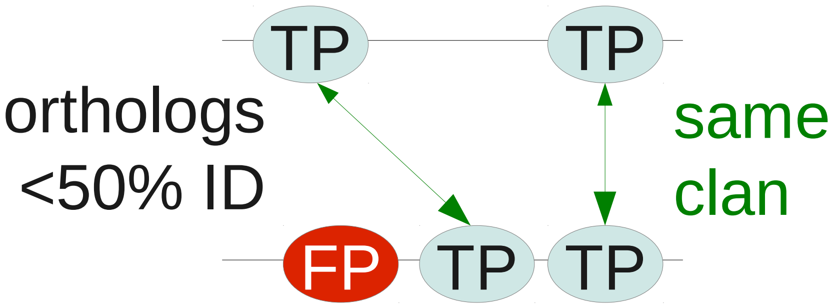
# Empirical null models

$$FDR = \#FP \,/\, ( \#FP + \#TP )$$

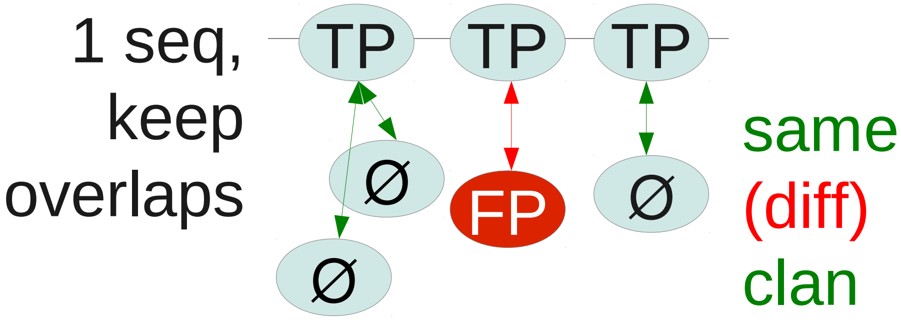
### 2nd order Markov Random


real
TP TP
+rand
FP
=
merged
TP FP TP

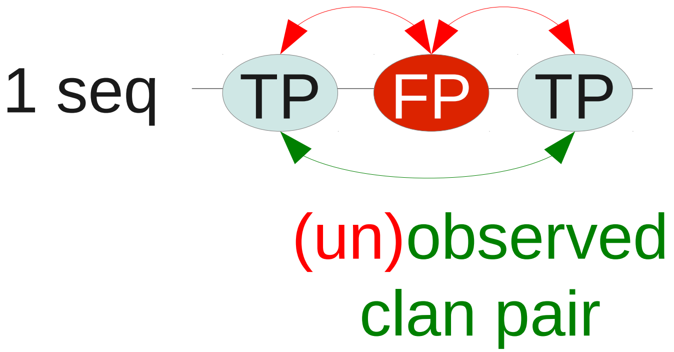
### Ortholog Set Coherence


TP TP
orthologs
<50% ID
same
clan
FP TP TP

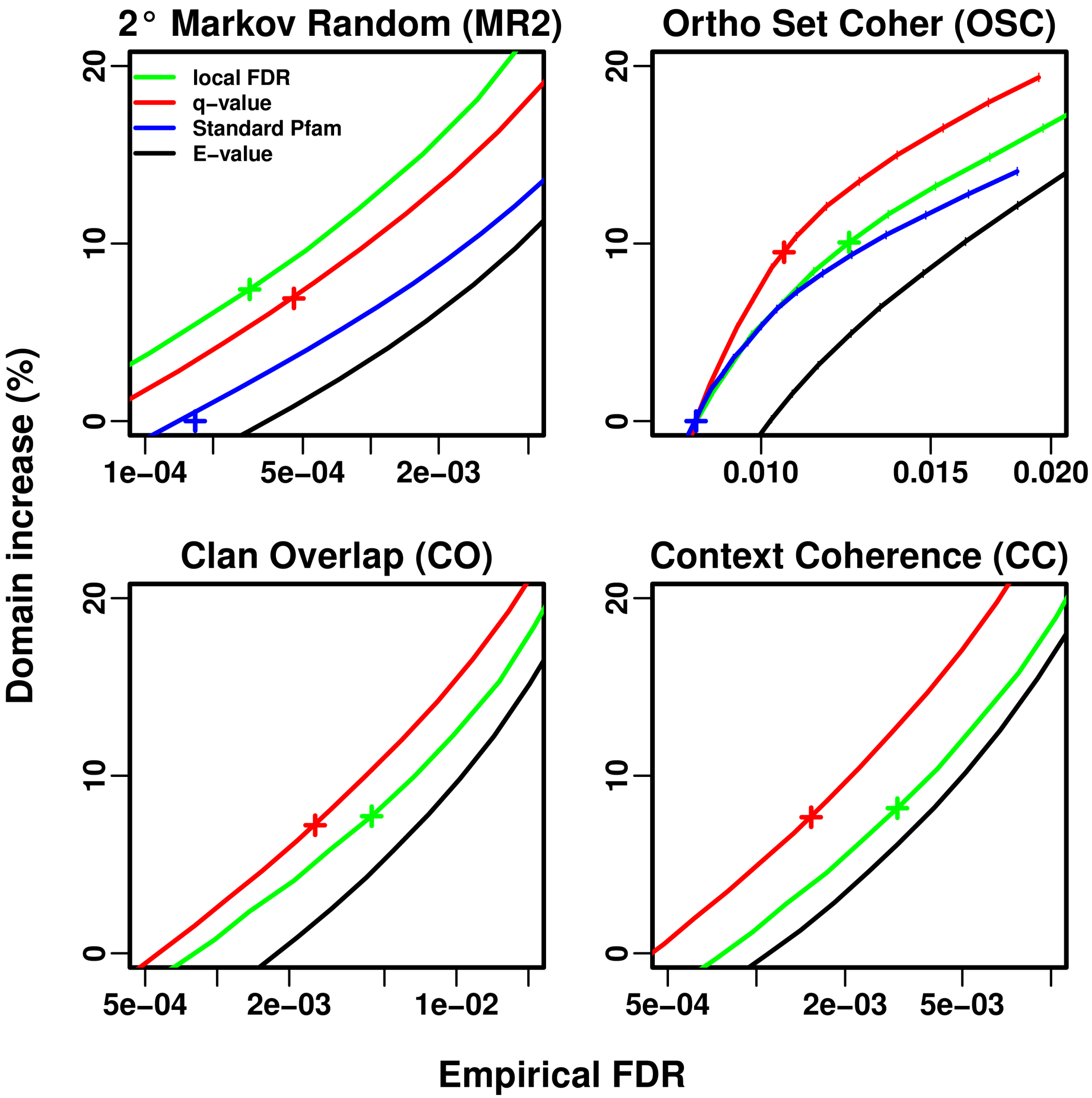
### Clan Overlap


1 seq,
keep
overlaps
TP TP TP
Ø
FP
Ø
Ø
same
(diff)
clan

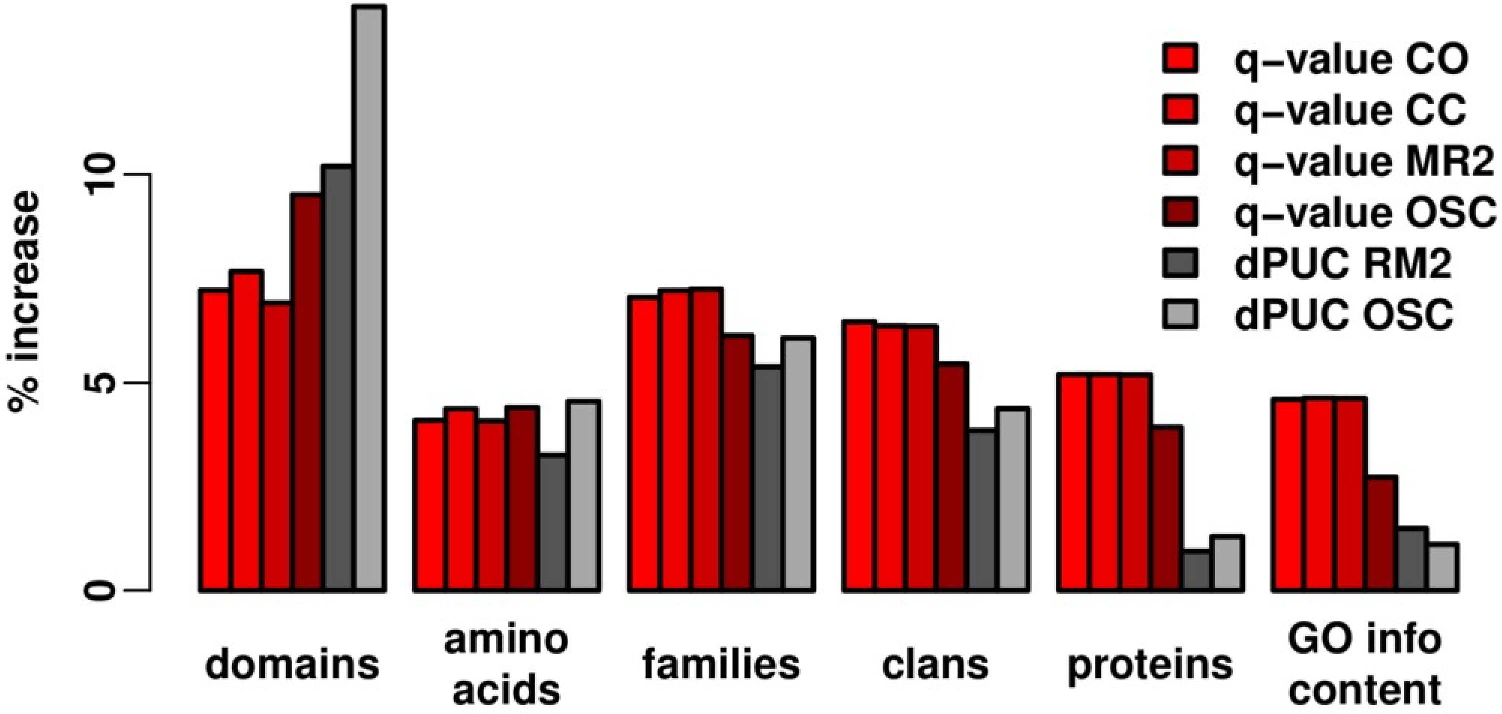
### Context Coherence


1 seq
TP FP TP
(un)observed
clan pair

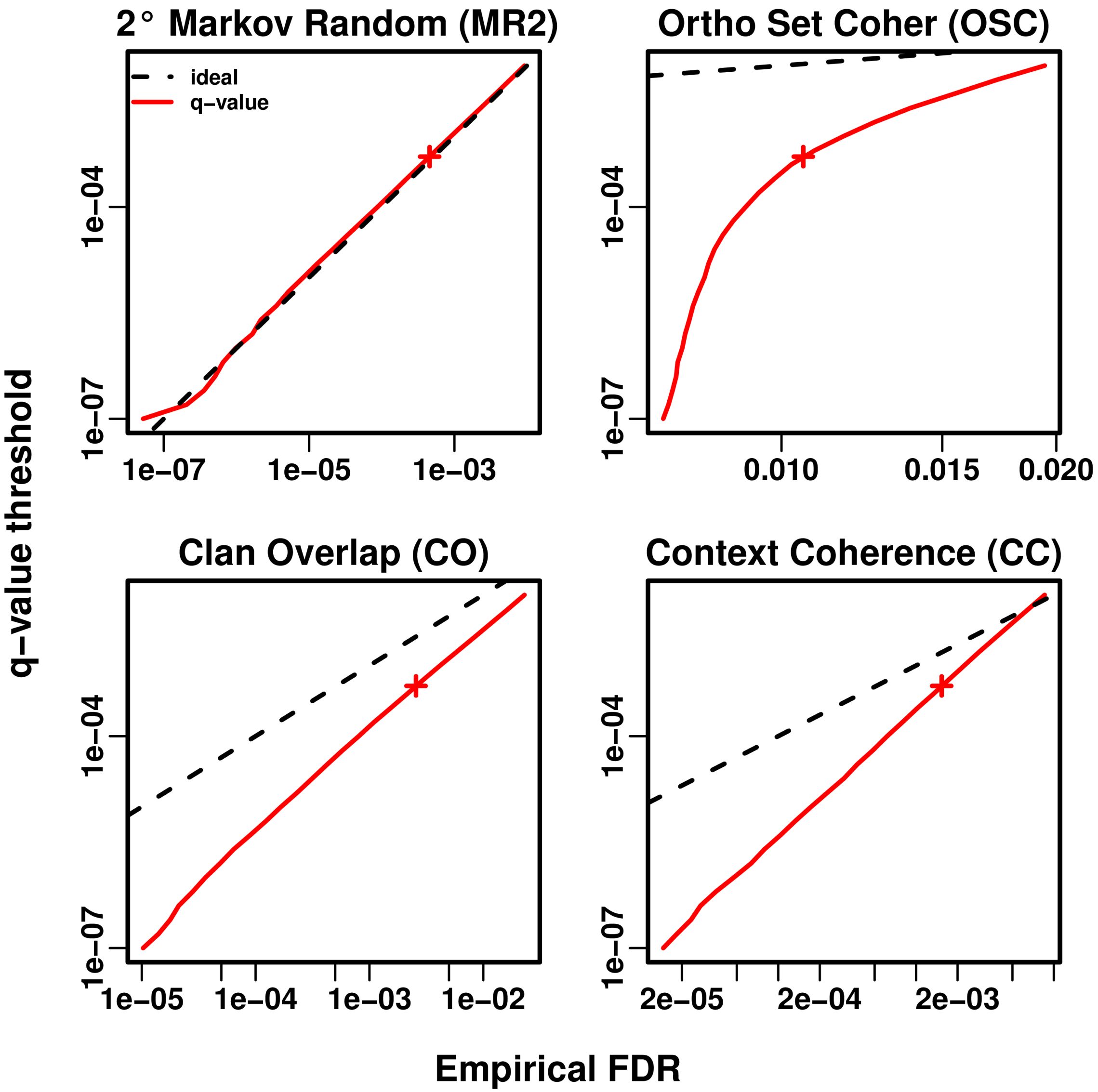
# *q > local FDR*


2° Markov Random (MR2)
Ortho Set Coher (OSC)
Clan Overlap (CO)
Context Coherence (CC)
local FDR
q–value
Standard Pfam
E–value
Domain increase (%)
Empirical FDR
0
10
20
1e–04
5e–04
2e–03
0.010
0.015
0.020
5e–04
2e–03
1e–02
5e–04
2e–03
5e–03

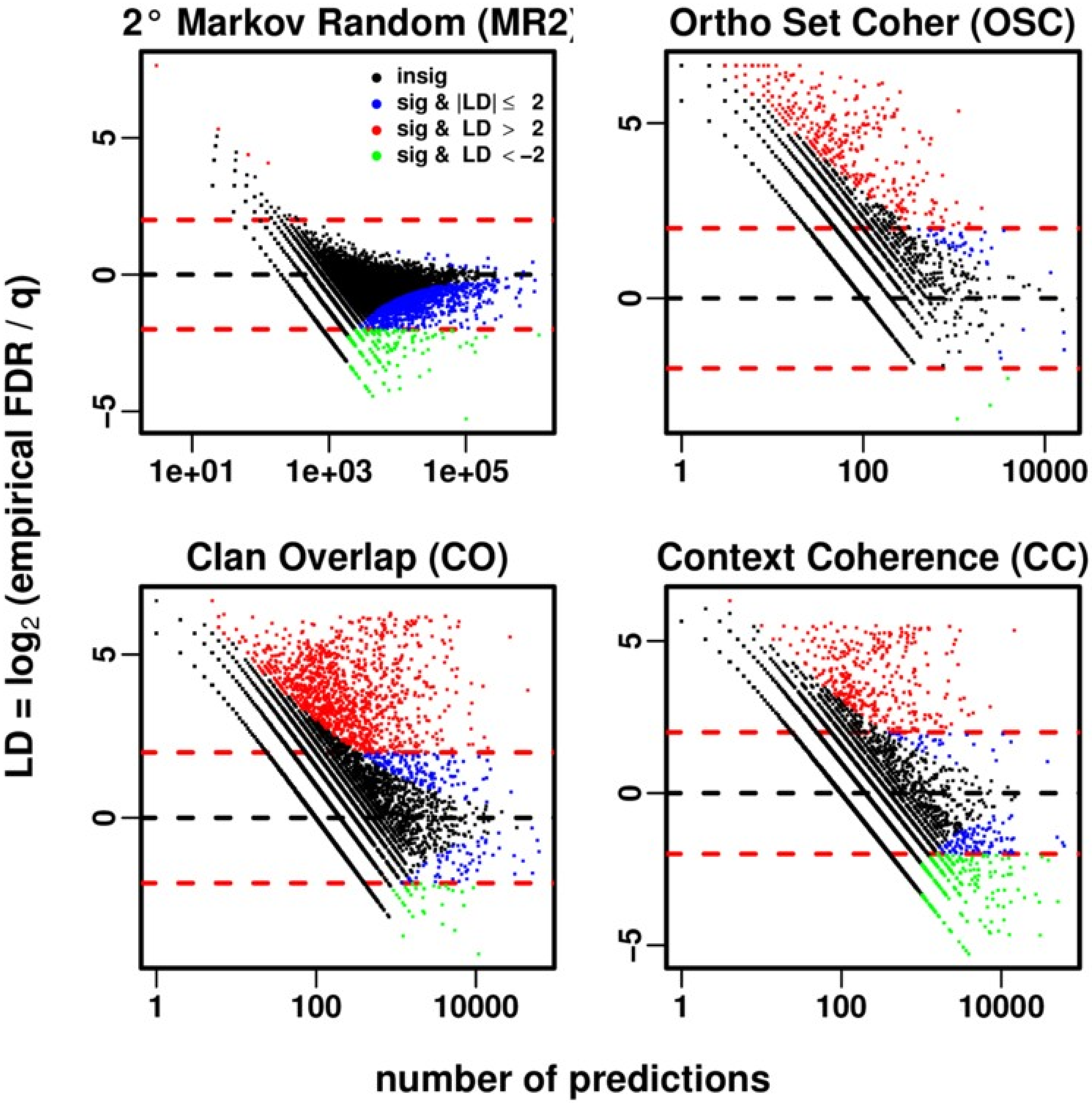
# Improving more than just domains



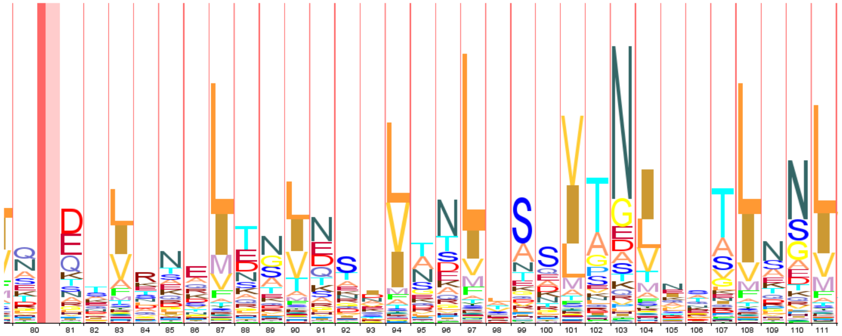
# The *q*-values underestimate empirical *FDRs*



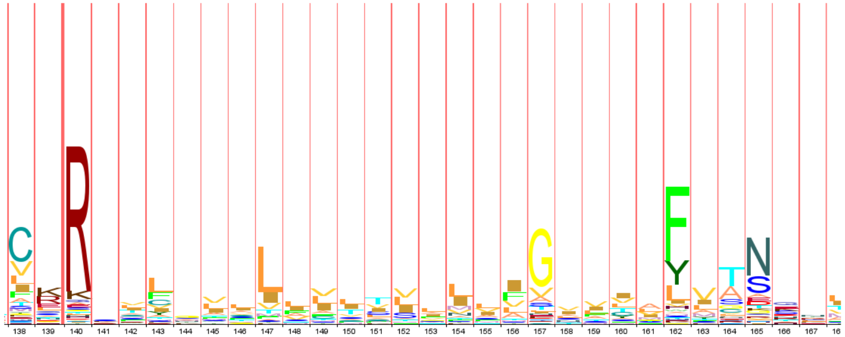
# Per family analysis of noise



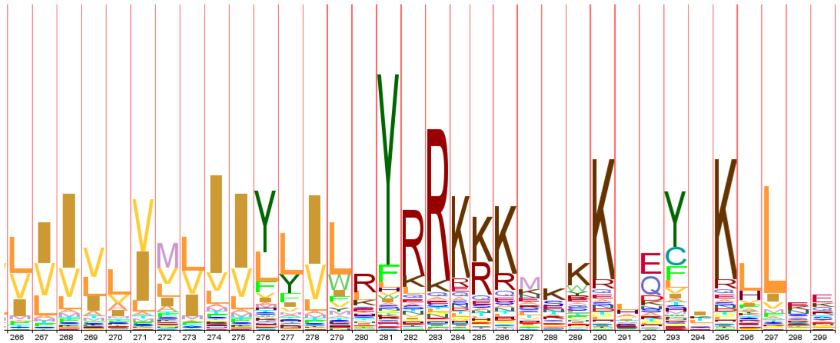
# Classes of noisy domains



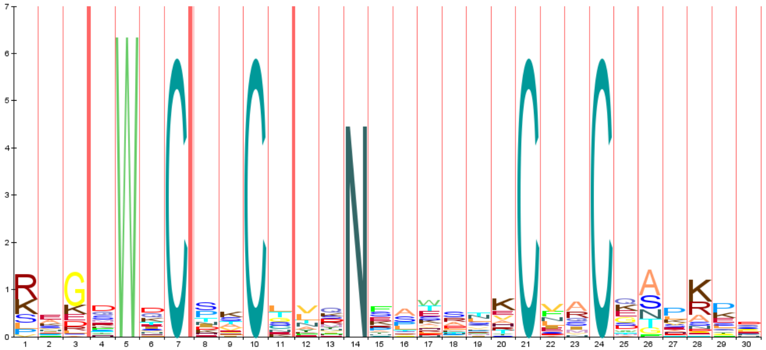
**Coiled coils: 403 (3%)**



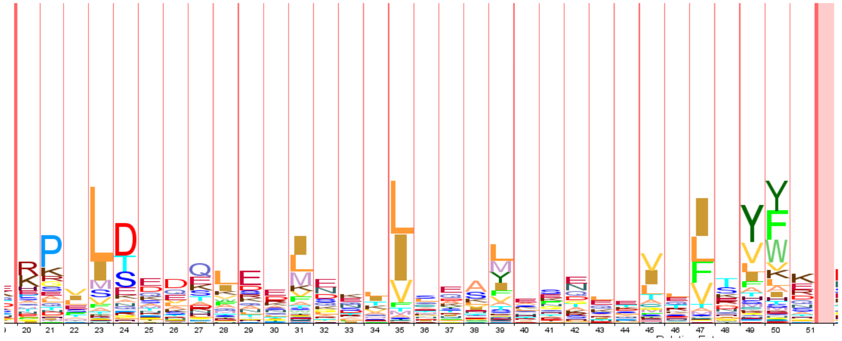
**Transmembranes: 1466 (12%)**



**Low complexity: 1987 (16%)**



**Cysteine-rich: 587 (5%)**

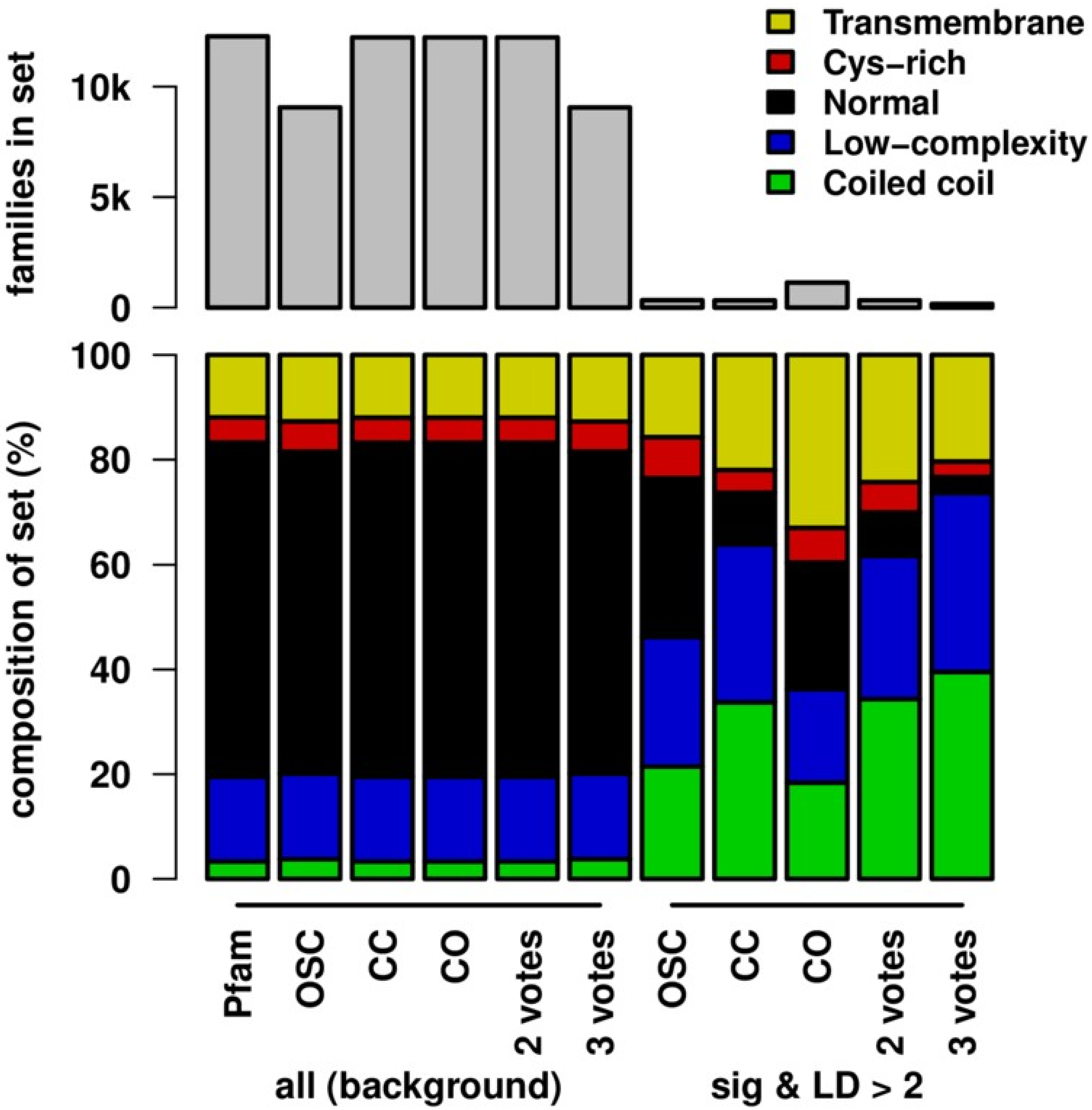


**Normal: 7830 (64%)**

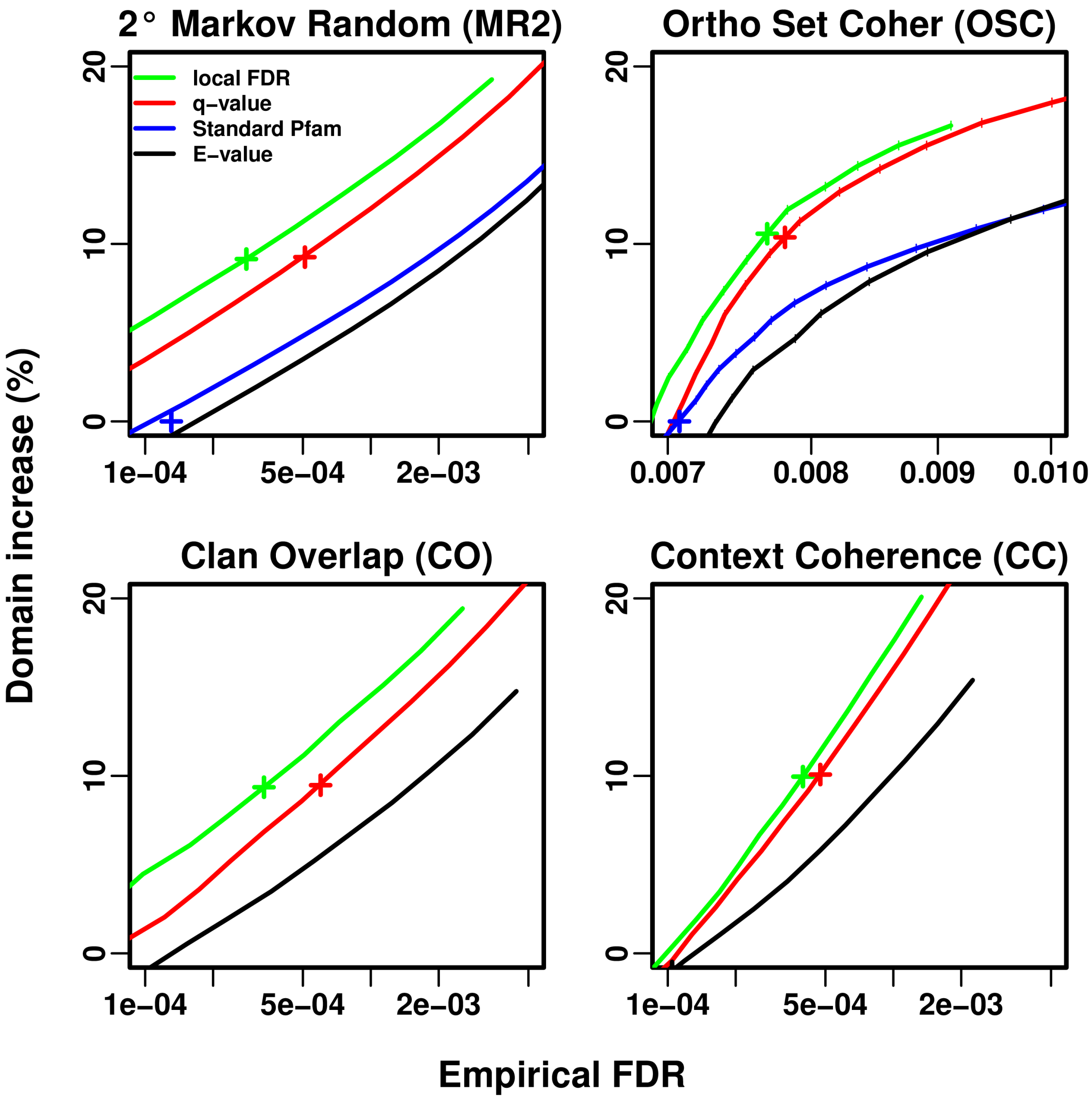
Similar but non-homologous families (different clans)

Rackham *et al*. JMB 403, 480–493 (2010).
Wong, *et al.* PLoS Comput Biol 6, e1000867 (2010).

# Classes enriched in noisy domains



# *Local FDR > q*
# in families with correct stats



# Conclusions

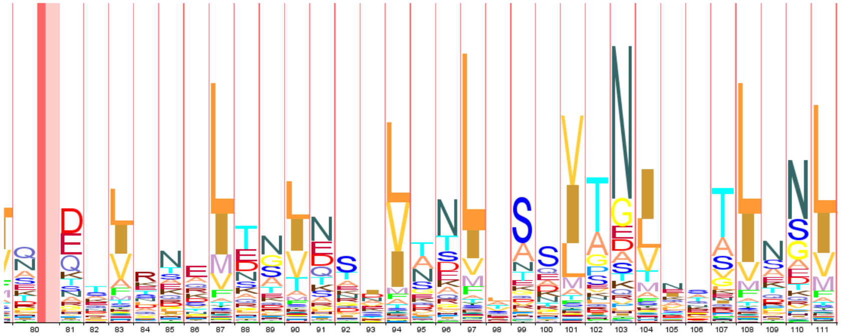
The *q*-value and *local FDR* are better for domains

- And likely better for regular sequence database searches, iterated searches, orthology prediction
- *E*-values do not control posterior error probability

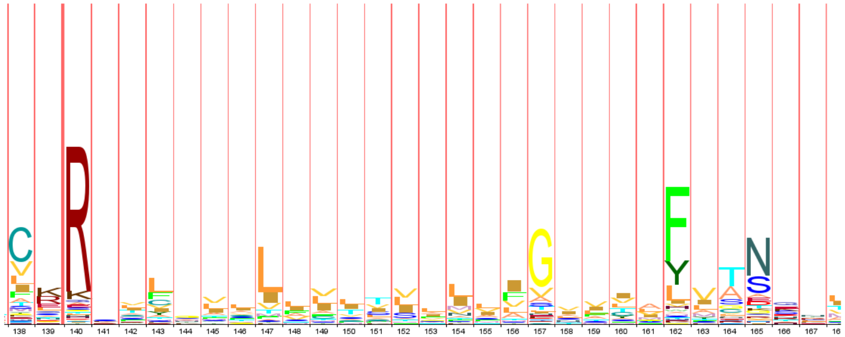
Presented novel empirical null models

- Needed to verify theory is correct
- Uses common-sense biological information and real, full protein sequences
- Structural benchmarks (i.e. SCOP) are limited to well-studied, single domains from model organisms, and exclude coiled coils and transmembrane domains

# Noisy domains



Coiled coils



Transmembranes

These remain large problems in all sequence analysis.

Solutions?

- Cannot prevent or ignore these queries
- Masking removes too much information
- Benchmarks not powerful enough to give better thresholds
- Can we properly handle these common, correlated patterns that do not imply homology?

Rackham *et al*. JMB 403, 480–493 (2010).
Wong, *et al.* PLoS Comput Biol 6, e1000867 (2010).

PRINCETON UNIVERSITY

- Mona Singh, Computer Science
  - Jesse Farnham
  - Dario Ghersi
  - Peng Jiang
  - Shilpa Nadimpalli
  - Anton Persikov
  - Yuri Pritykin
  - Pawel Przytycki
  - Josh Wetzel
- Thesis Committee
  - John Storey
  - Leonid Krugliak
  - Saeed Tavazoie

- Manuel Llinás, Molecular Biology
  - Lindsey Altenhofen
  - Katie Baska
  - Simon Cobbold
  - Valerie Crowley
  - Björn Kafsack
  - Ian Lewis
  - Jessica O’Hara
  - Heather Painter
  - Joana Santos
  - Ariel Schieler
  - April Williams
- NSF GRFP

http://compbio.cs.princeton.edu/dpuc/