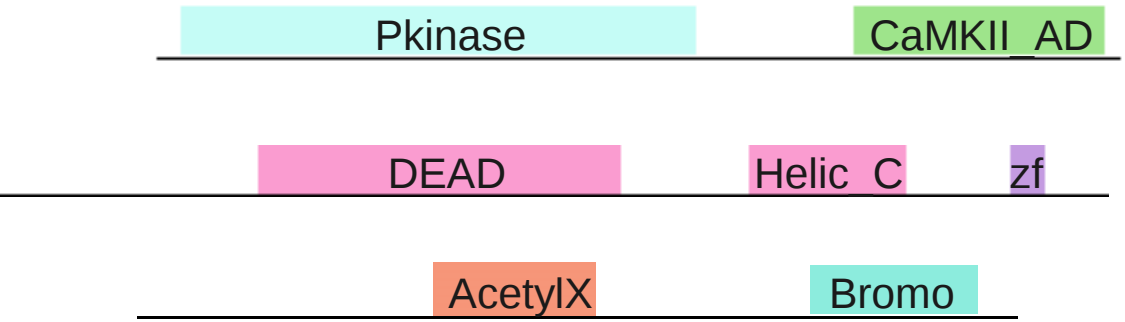
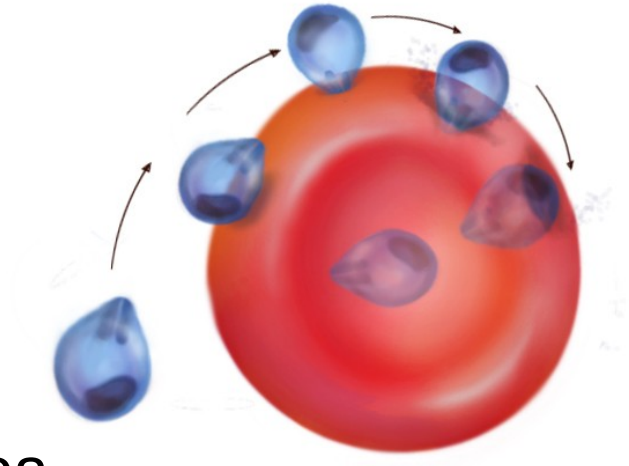
Protein domain prediction using context statistics, the false discovery rate, and comparative genomics, with application to *Plasmodium falciparum*



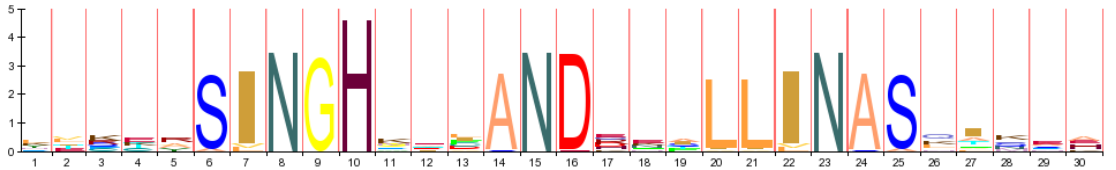


Alejandro Ochoa

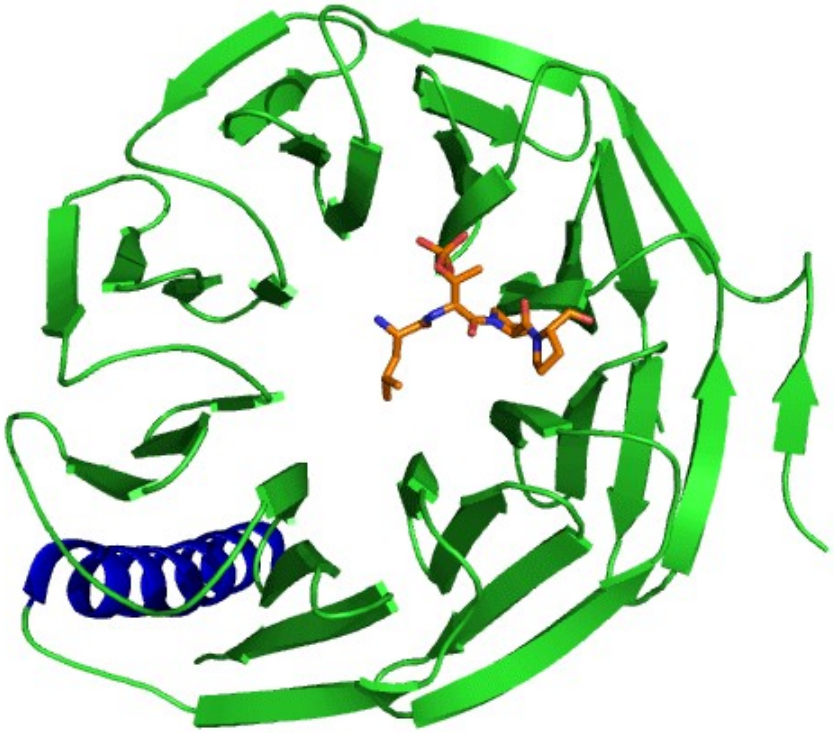
Molecular Biology, Princeton University

Final Public Oral, 2013-06-27

PRINCETON UNIVERSITY



Protein domains



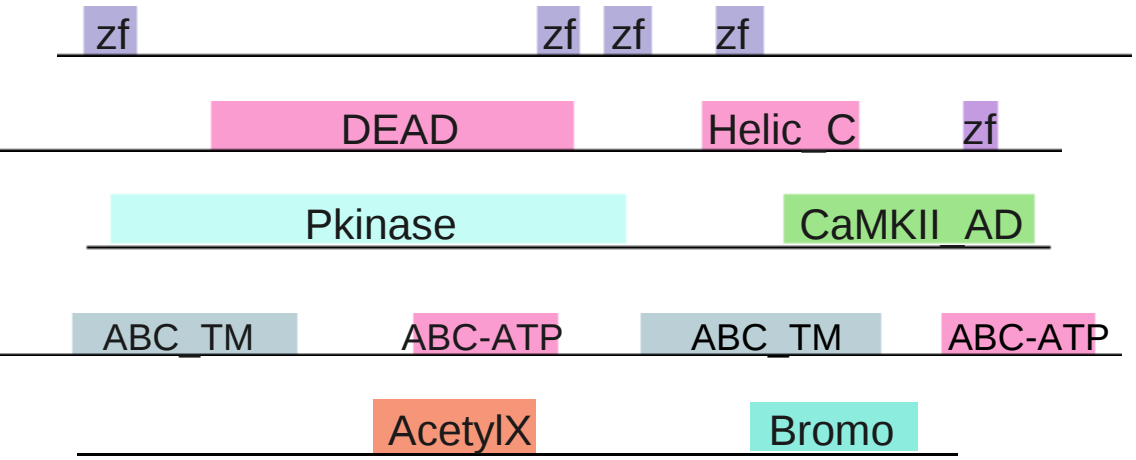
Structure

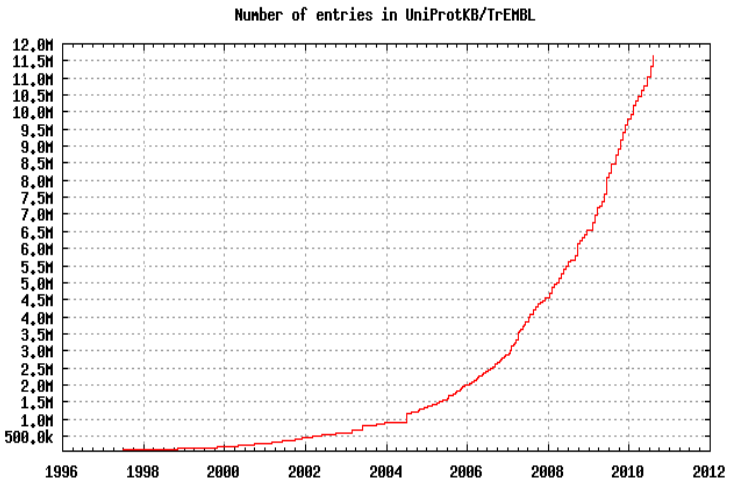
Evolution

Function

Sequence-based domain prediction:

Why predict domains?





For new sequences, before experiments start…

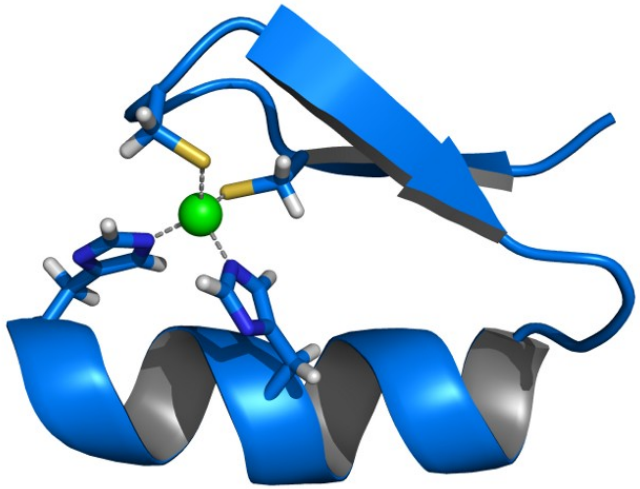
Domains may imply functions

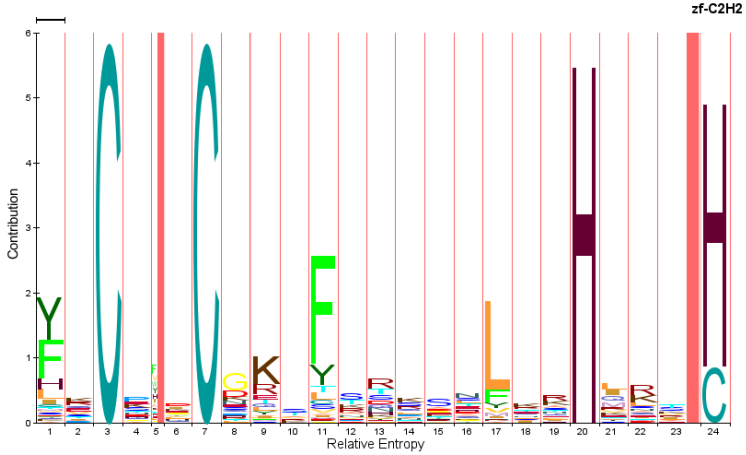
Experimental alternatives are unfeasible as protein databases grow exponentially

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
ZO20_XENLA/174-196      FMCA...DCG...RCFSVS...SSLKYHQRI...C
EVI1_HUMAN/217-239      IKCK...DCG...QMFSTT...SSLNKHRRF...C
ZO2_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
```

. . . . . .

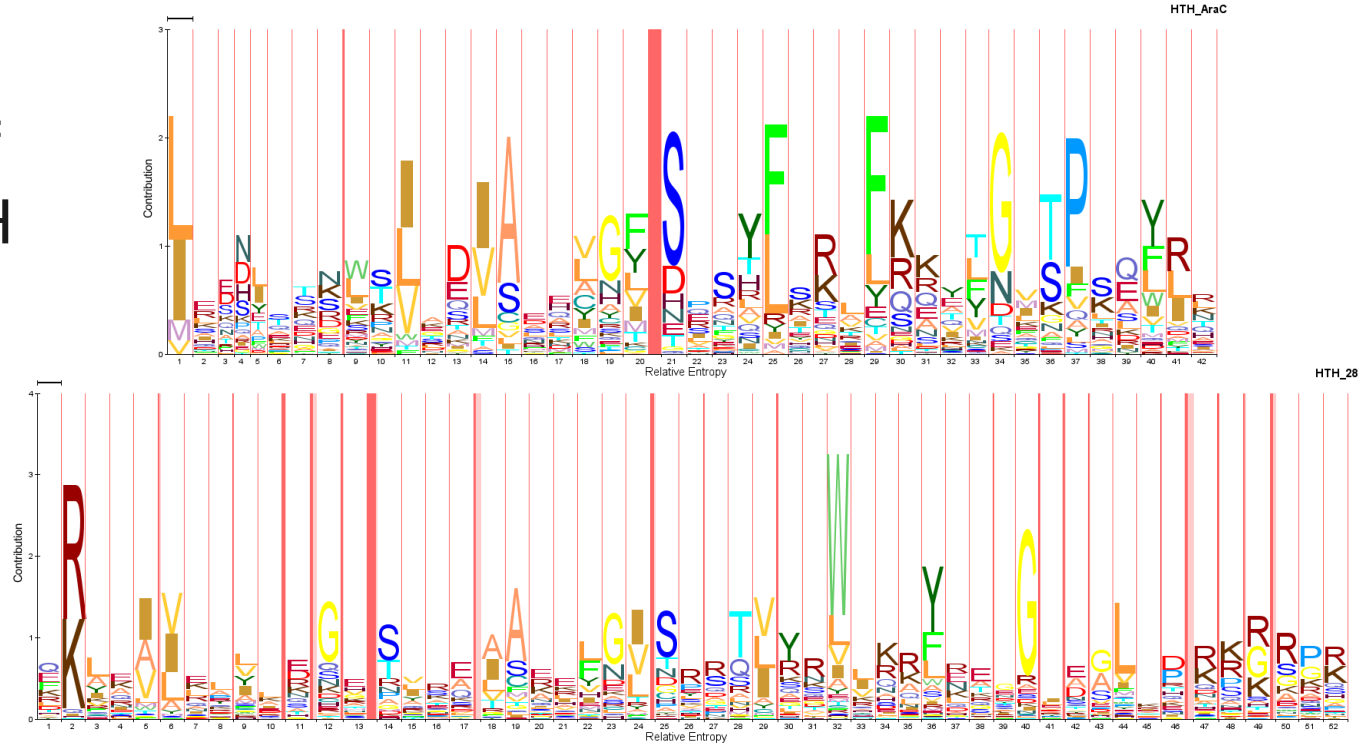




Databases of Domain Families

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

Two members of Pfam “Clan” HTH



Plasmodium falciparum

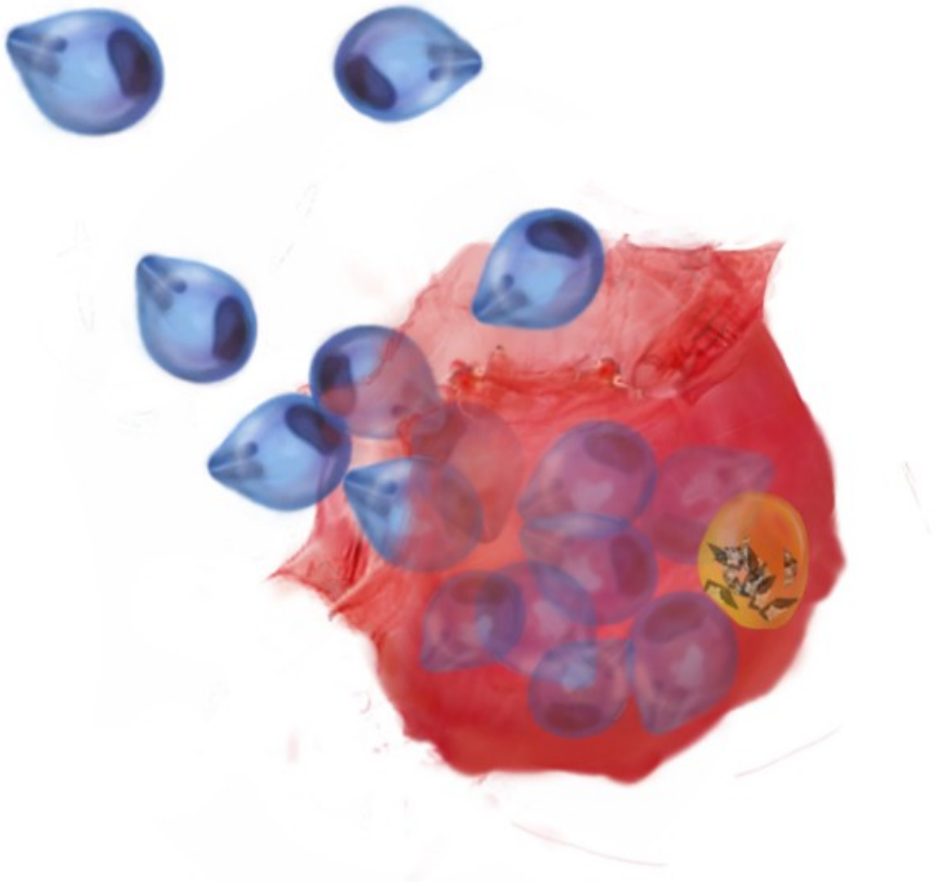
Malaria

Information challenges

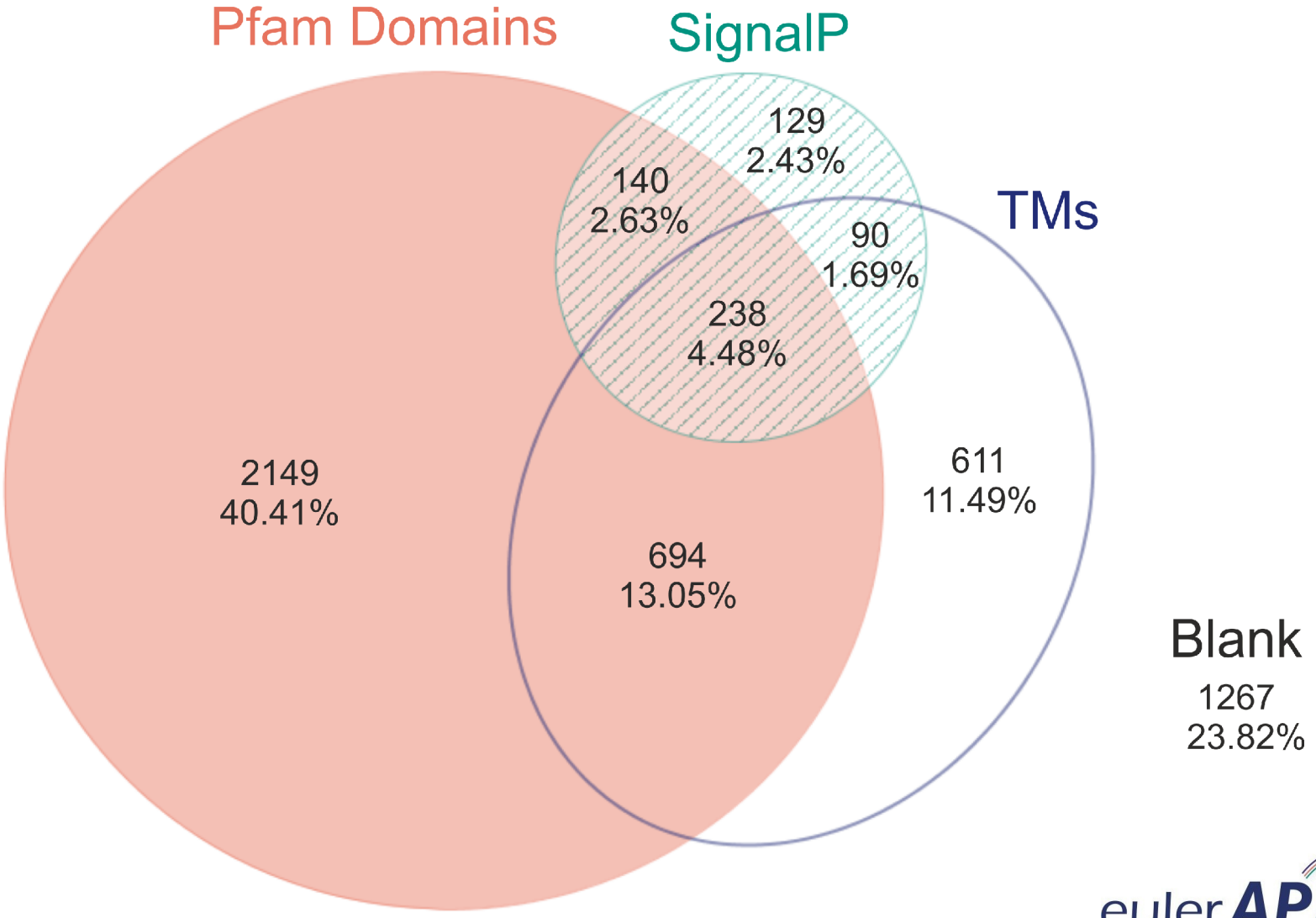
- Diverged eukaryote
- 80% AT-bias
- Low-complexity regions

Annotation

- 5.5K proteins
- 45% unknown function
  - 20% unknown in yeast
- 88% of annotations are bioinformatical

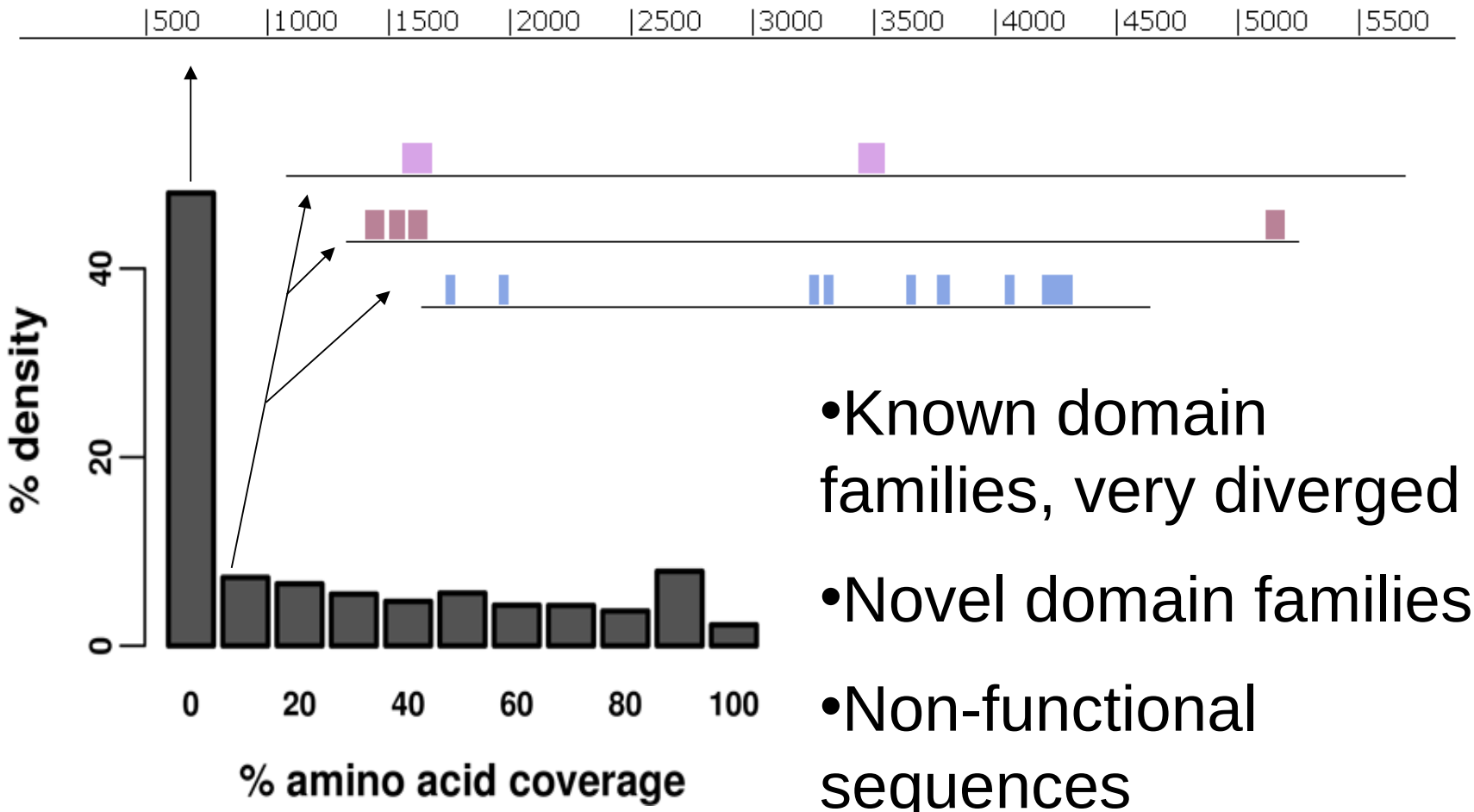


Sequence-based annotation of *Plasmodium falciparum* proteome



eulerAPE

Poor domain coverage of *Plasmodium falciparum*

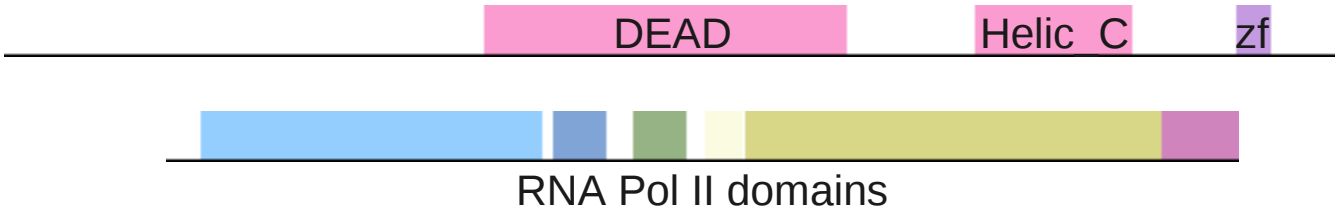


- Known domain families, very diverged
- Novel domain families
- Non-functional sequences

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
- Application and case studies in *P. falciparum*

Domain Prediction Using Context: dPUC



Background

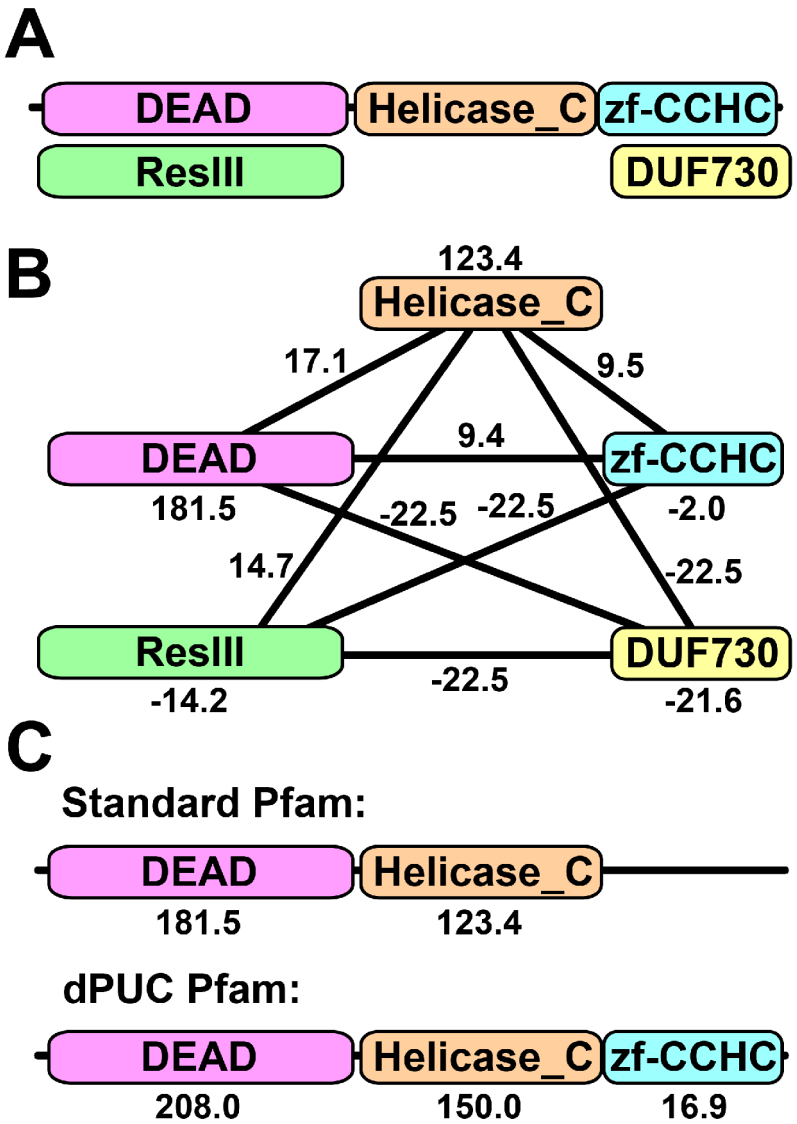
- Domains co-occur in limited combinations
- Domains are scored independently of each other

Idea

- Score domains in combination
- Context + Sequence evidence

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

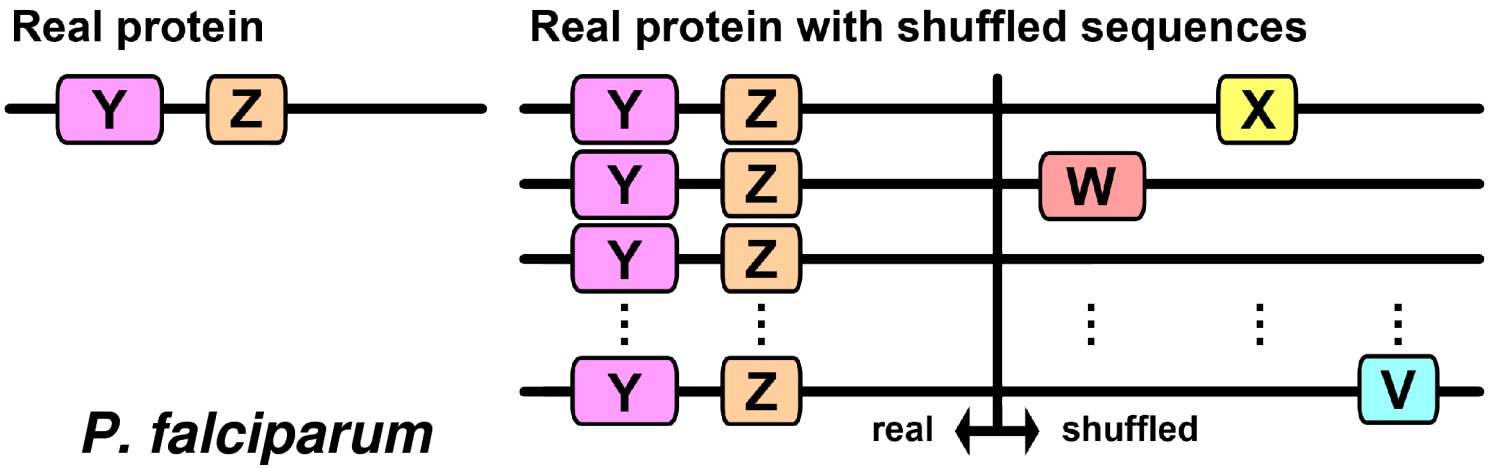
The dPUC method

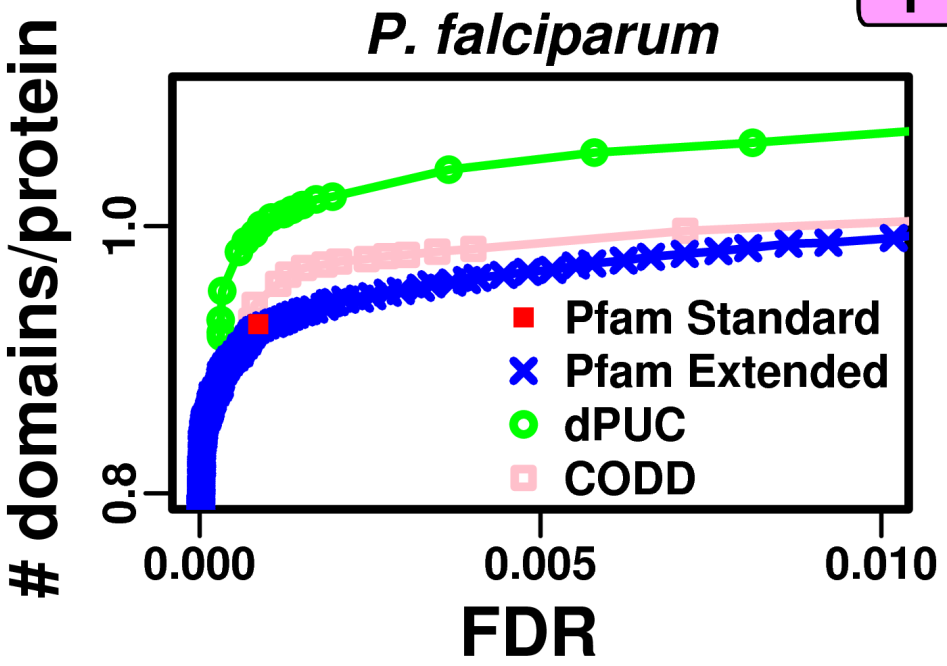


http://compbio.cs.princeton.edu/dpuc/

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

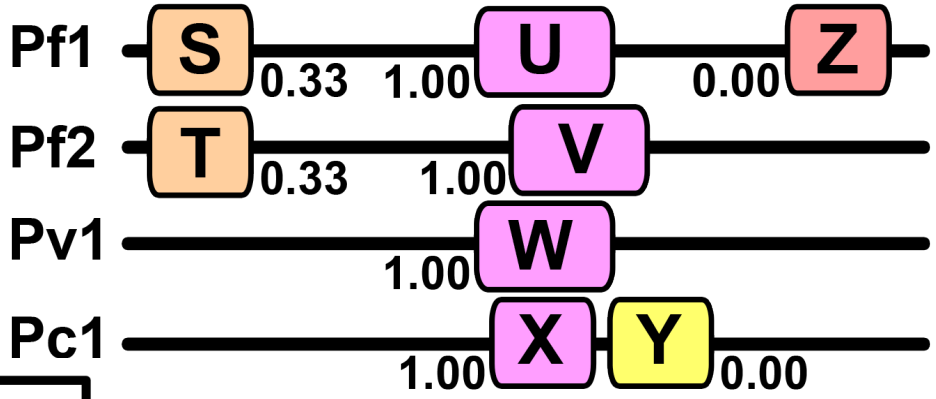
Improved signal to noise

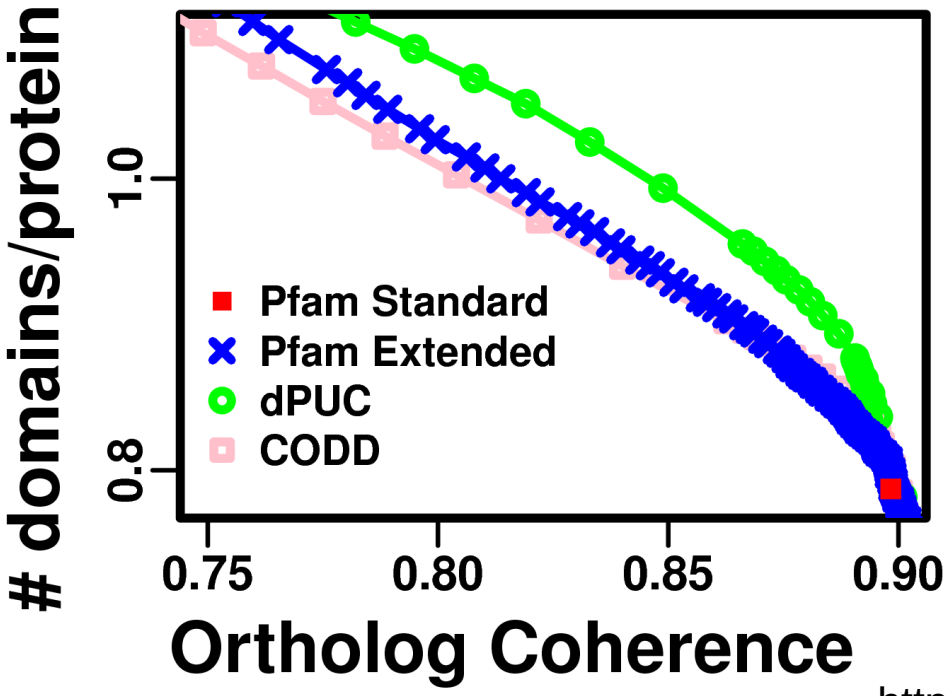




http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

Improved ortholog coherence on *Plasmodium* species





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

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

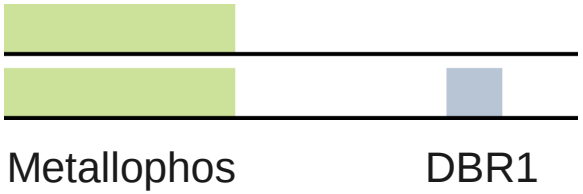
New predictions

Phosphatase -> RNA lariat debranching enzyme

P. falciparum

Std Pfam

dPUC



S. cerevisiae

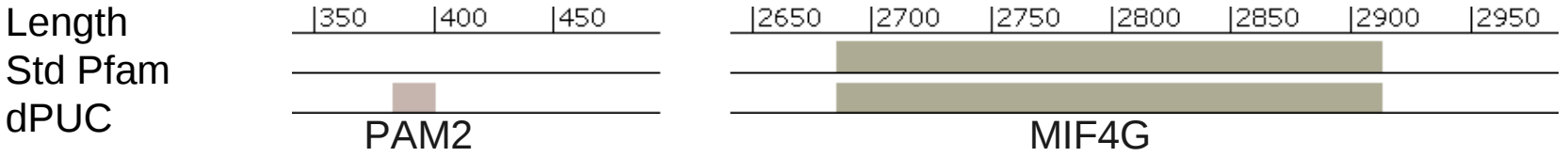
Std Pfam & dPUC

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

P. falciparum



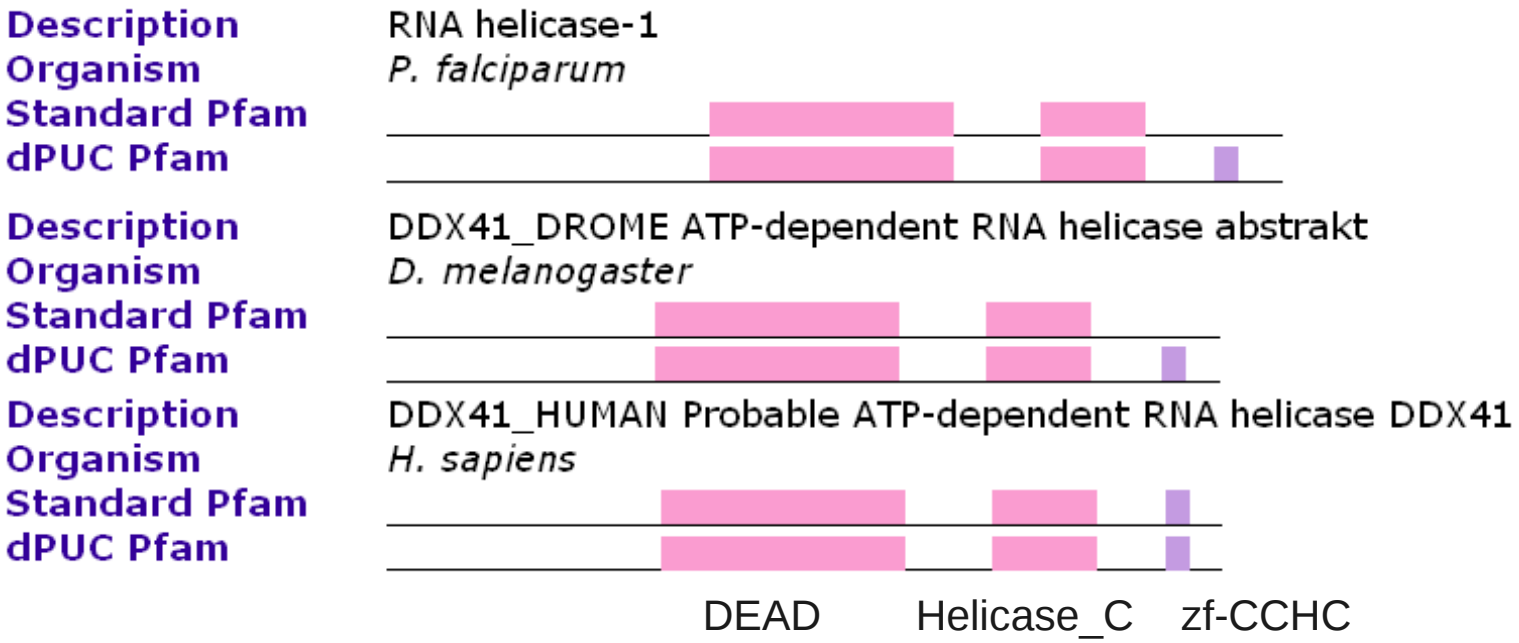
H. sapiens

Interested malaria curator:
Hagai Ginsburg,
Hebrew U of Jerusalem

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

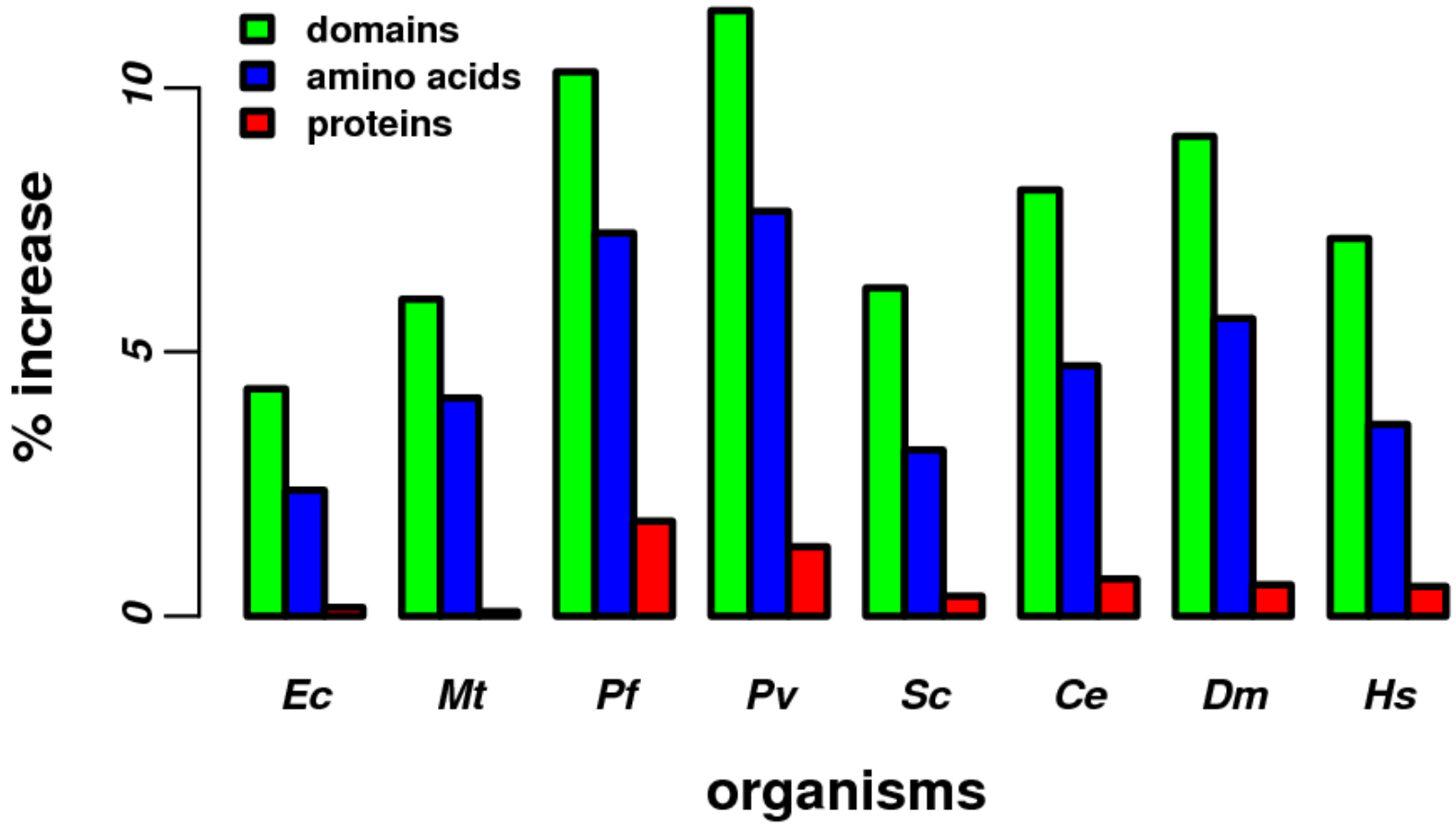
New predictions

RNA helicase -> mRNA sequestration



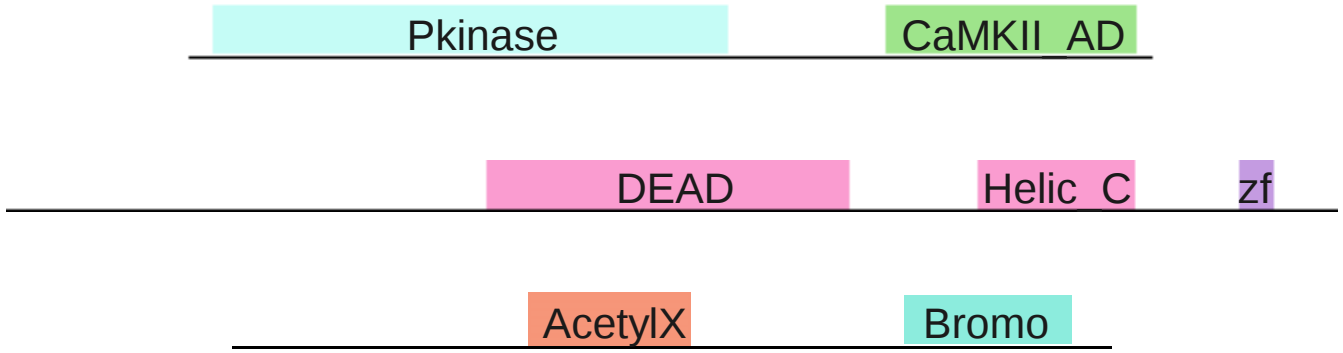
http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

dPUC increases coverage



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

Domain context



Complements sequence evidence

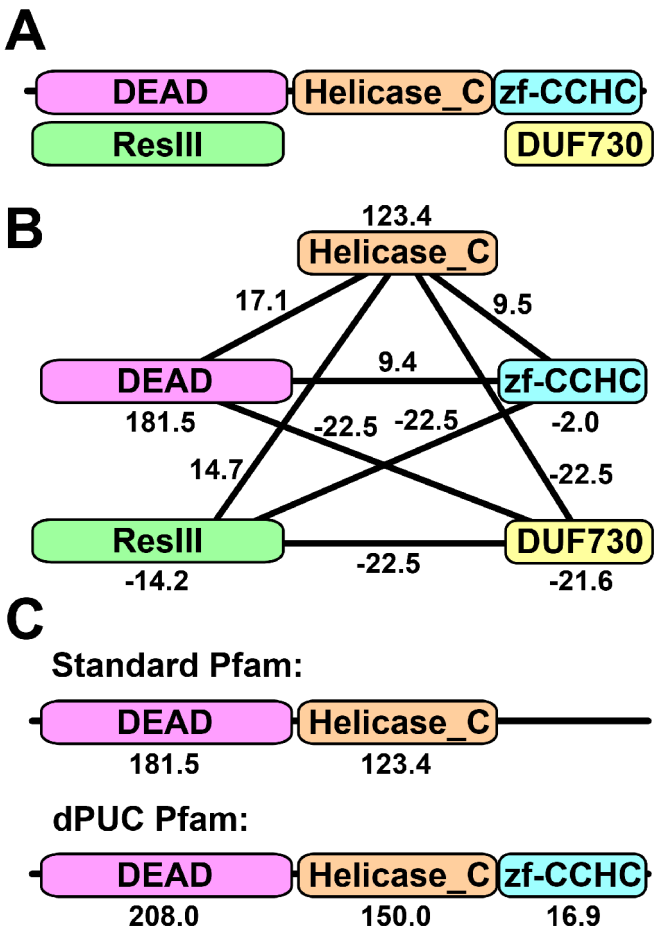
Improves domain predictions

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?

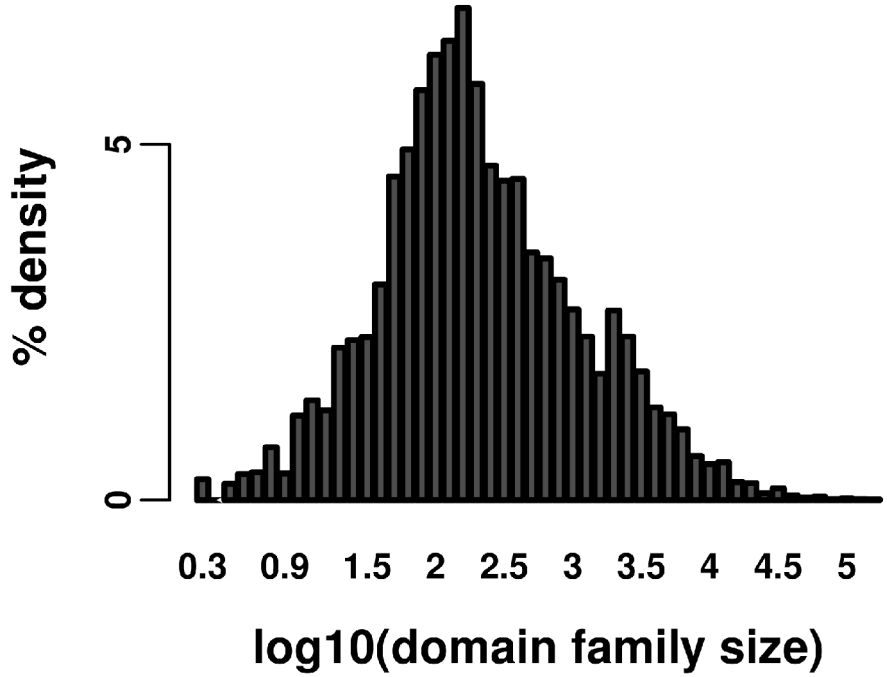
http://compbio.cs.princeton.edu/dpuc/
Ochoa, *et al*. BMC Bioinformatics 12, 90 (2011).

Outline of results

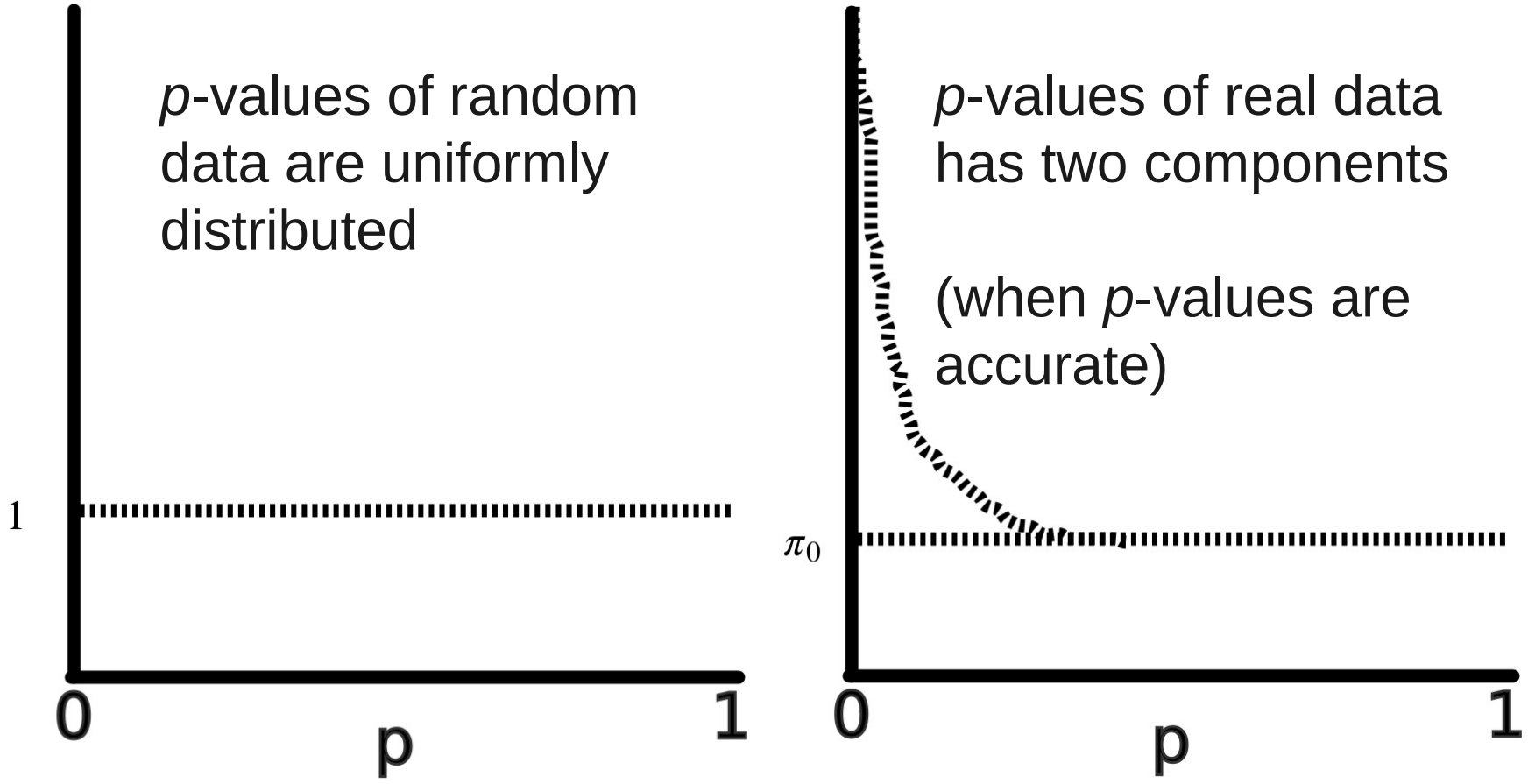
- Domain prediction using context
  - Application to malaria parasite
- Optimal *FDR* control for domains
  - Family-based *q*-values, adapted to domain problem
  - *Local FDR* optimizes problem, but *q*-values are more robust
  - Large problems with coiled coils, transmembrane domains
  - Implications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

FDR = E/n

FDR = average posterior error probability

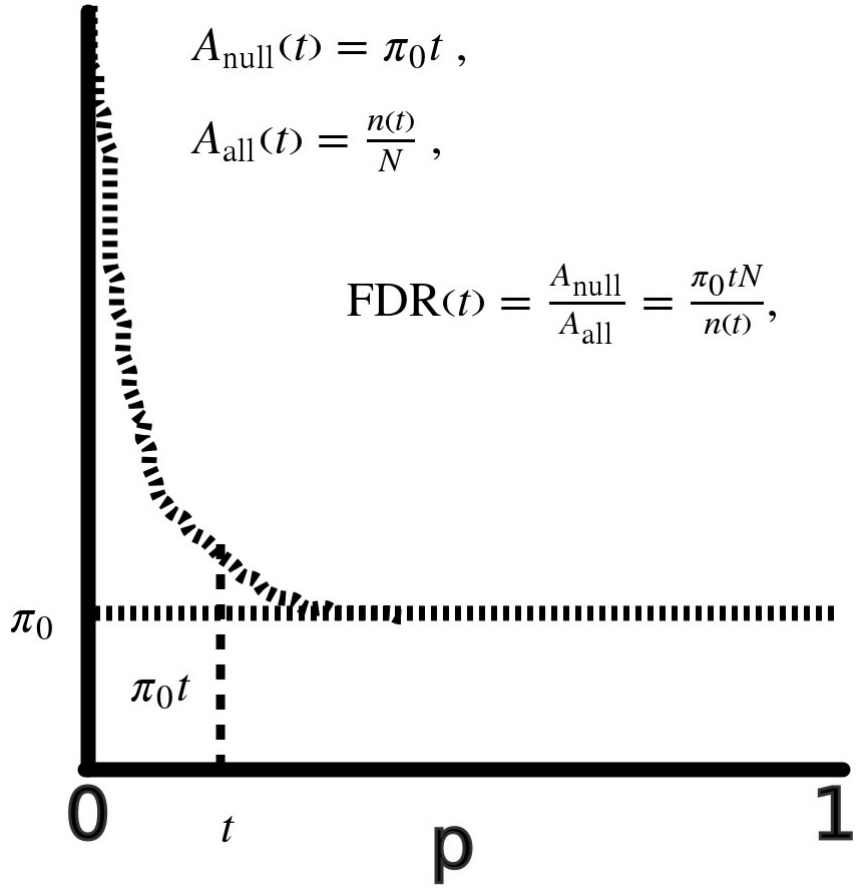


Computing *q*-values



Step 1: estimate π_0 (proportion of data that is false)

Computing *q*-values



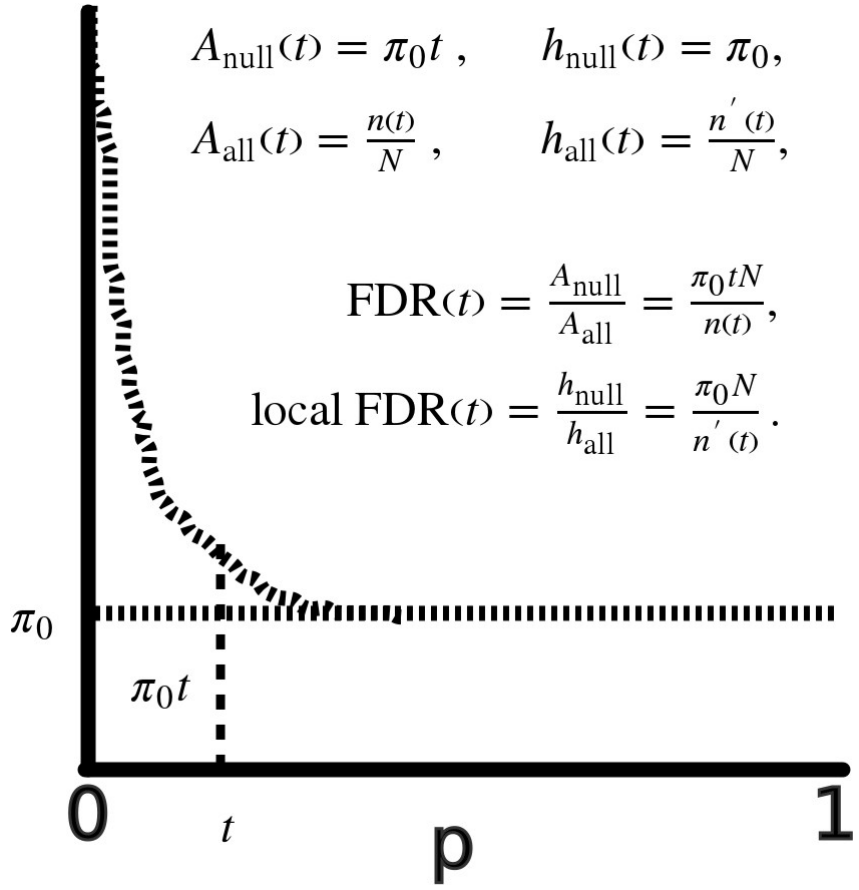
Step 2: Directly estimate *FDR*(*t*) for all thresholds *t*

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

Step 3: Ensure monotonicity

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

FDR and *local FDR*



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

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

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

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

Equal family *local FDR*s solve this optimization!
Also holds when constraining combined *E*-value!

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

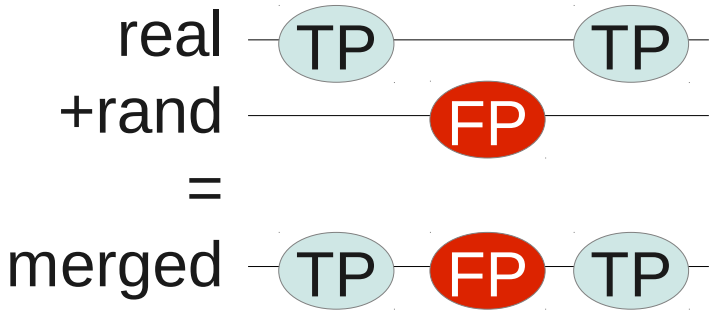
- **FP if domain doesn't co-occur with any stronger domains**

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

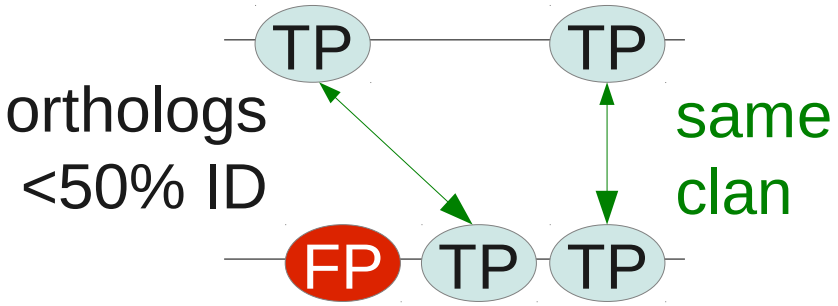
Empirical null models

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

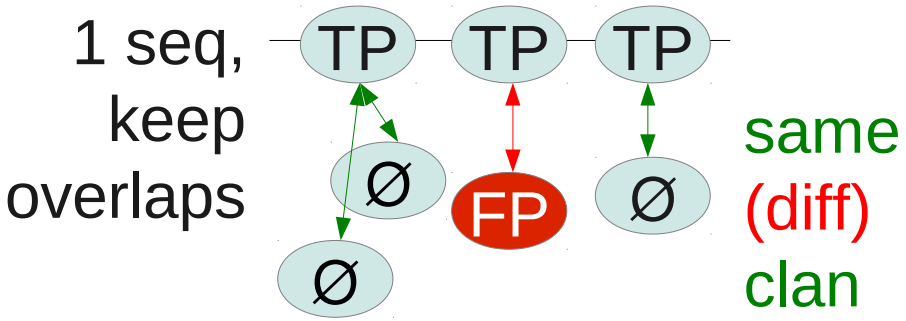
2nd order Markov Random



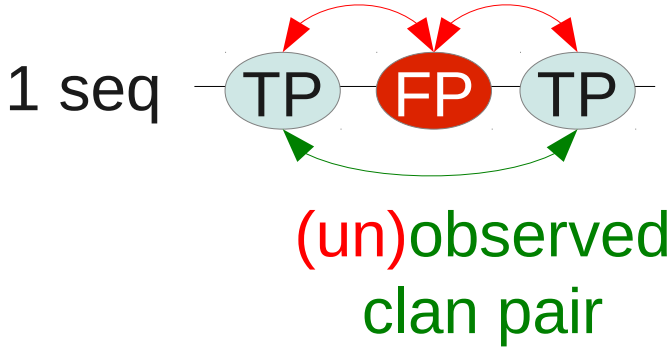
Ortholog Set Coherence



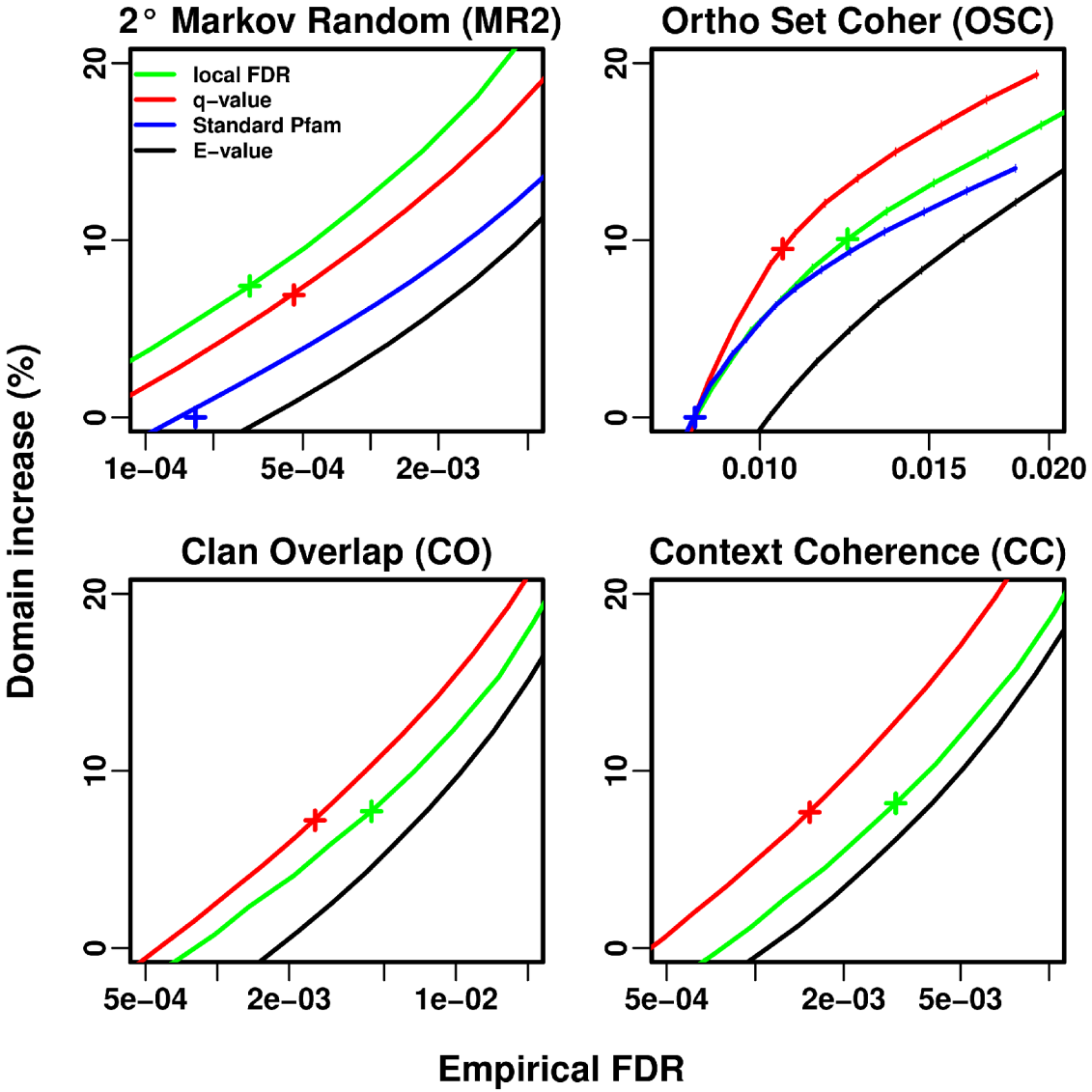
Clan Overlap



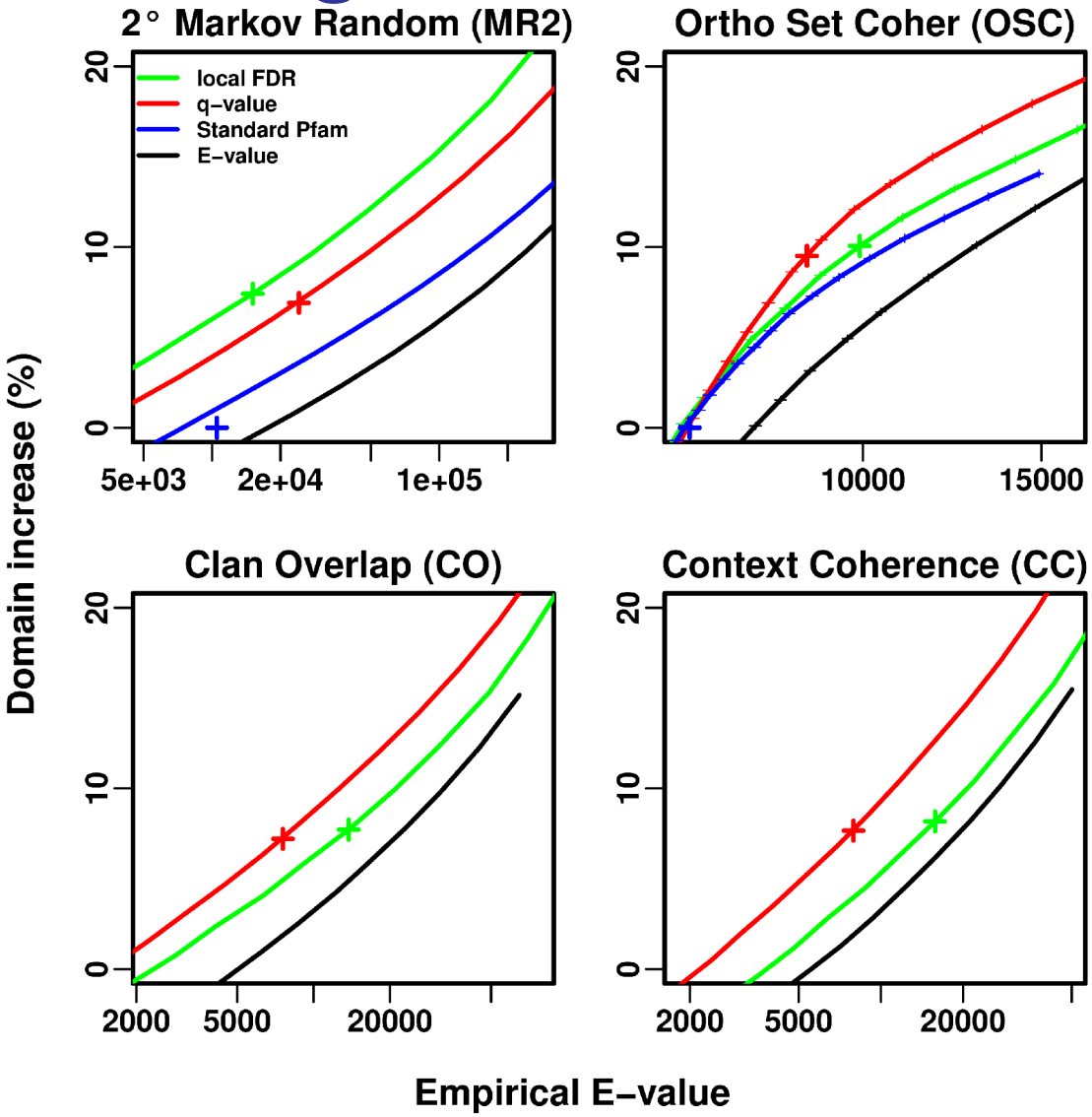
Context Coherence



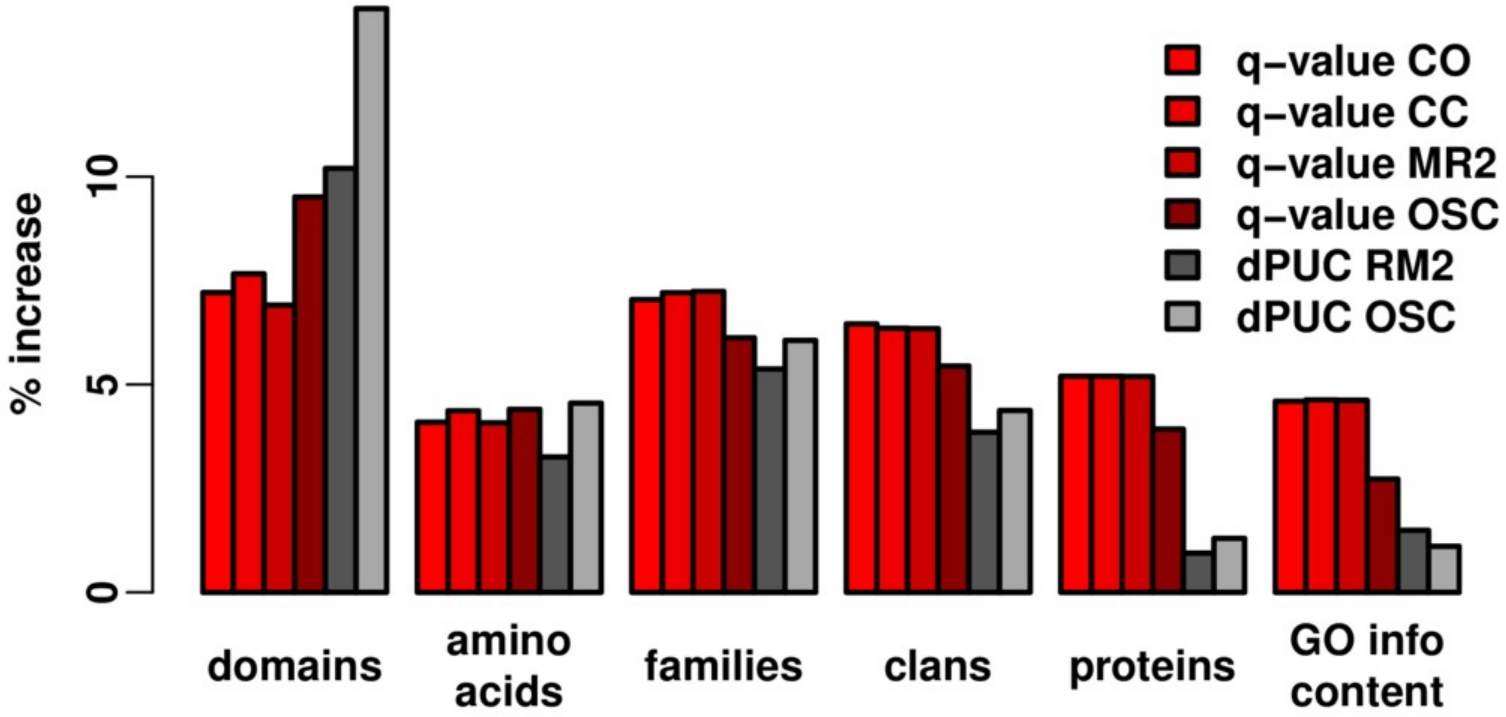
q > local FDR


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−03

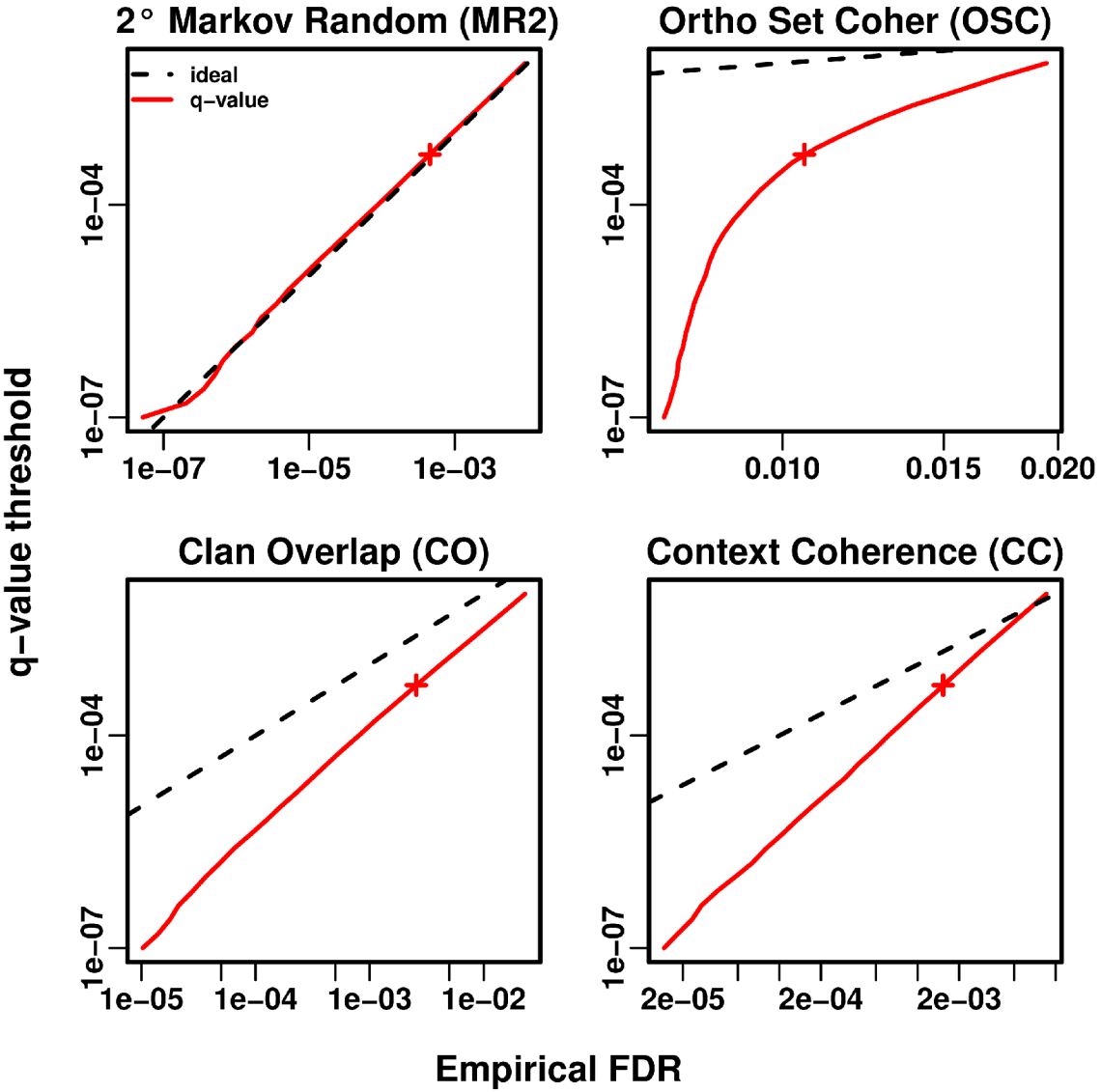
q > *local FDR*, *E* worst at constraining combined *E*-value!



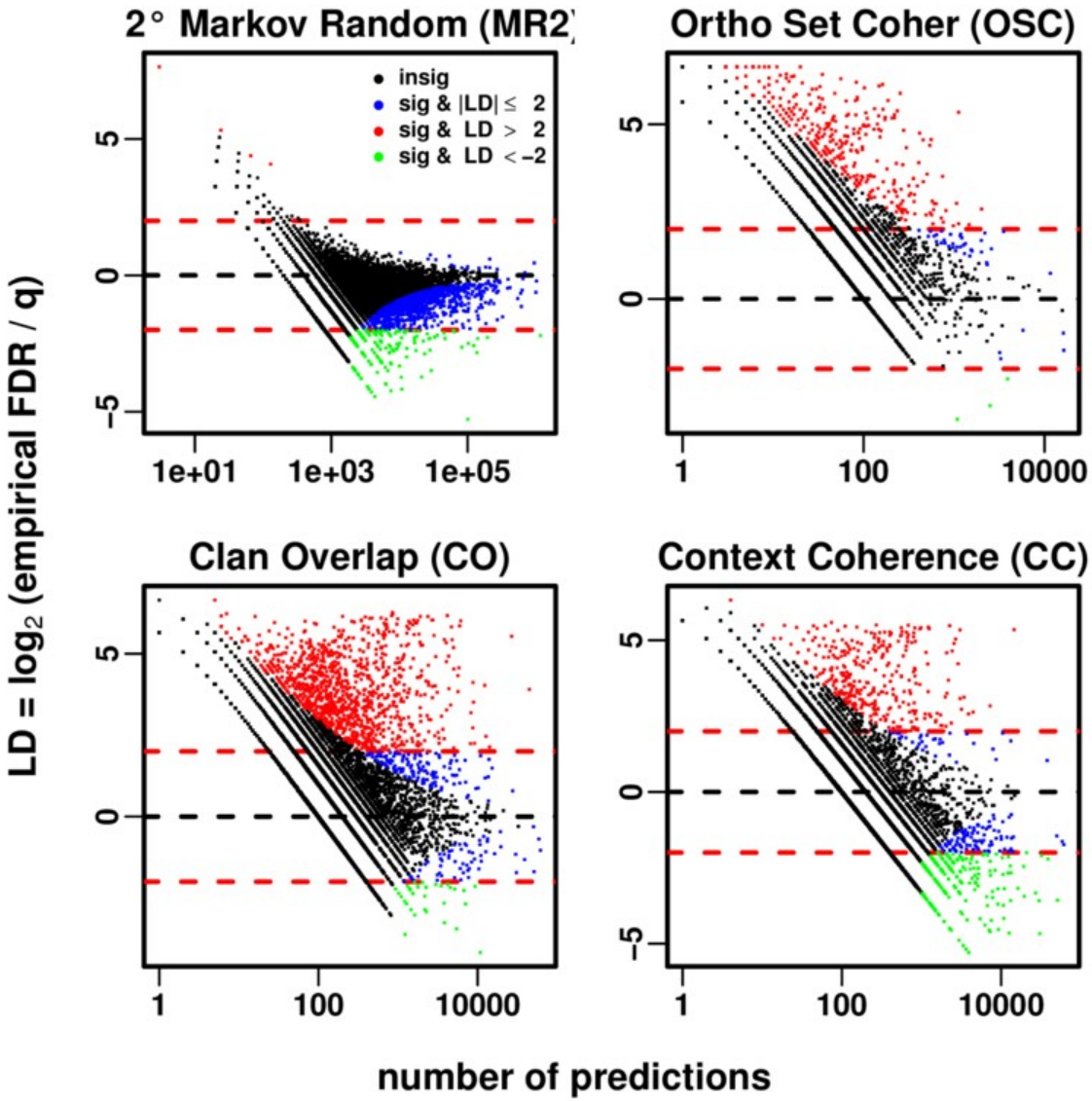
Improving more than just domain counts



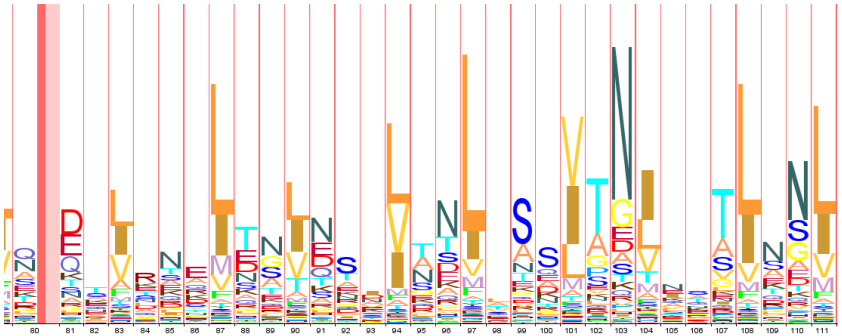
The *q*-values underestimate empirical *FDR*s



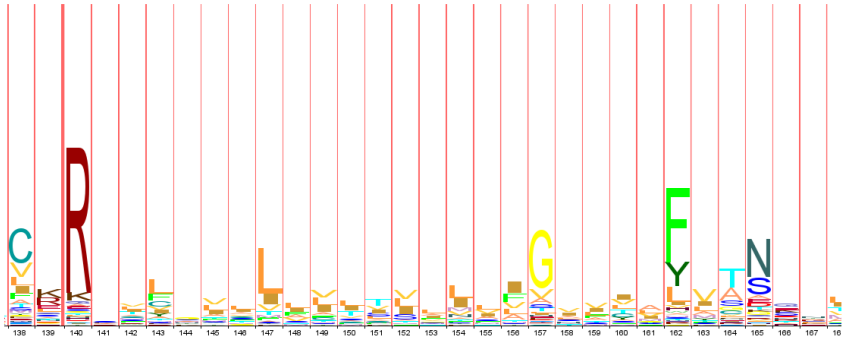
Per family analysis of noise



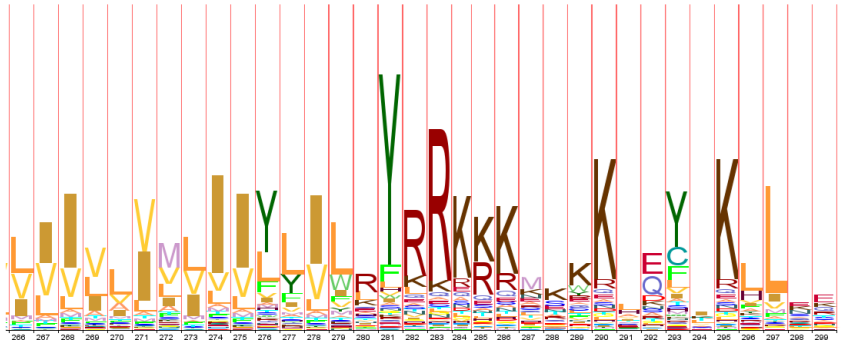
Classes of noisy domains



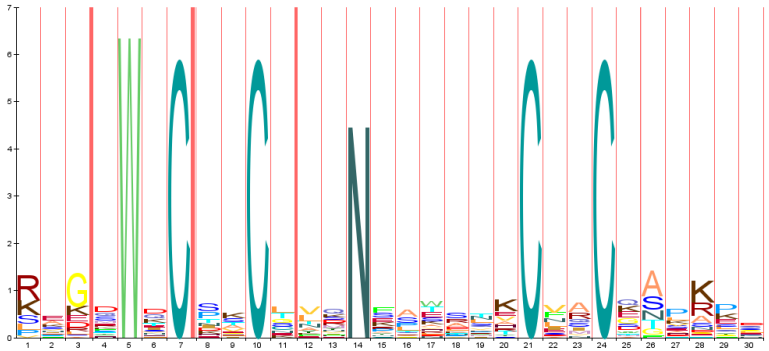
Coiled coils: 403 (3%)



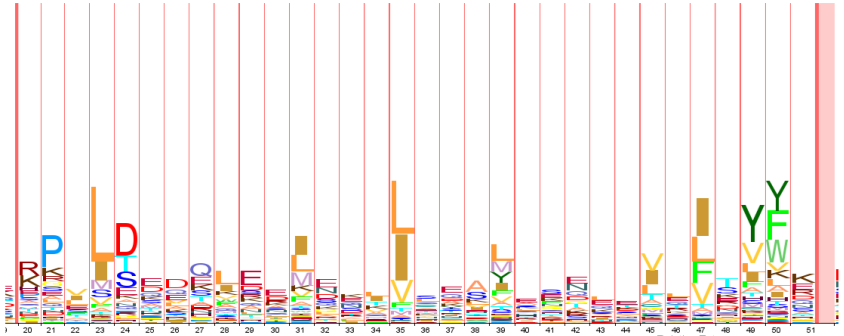
Transmembranes: 1466 (12%)



Low complexity: 1987 (16%)



Cysteine-rich: 587 (5%)

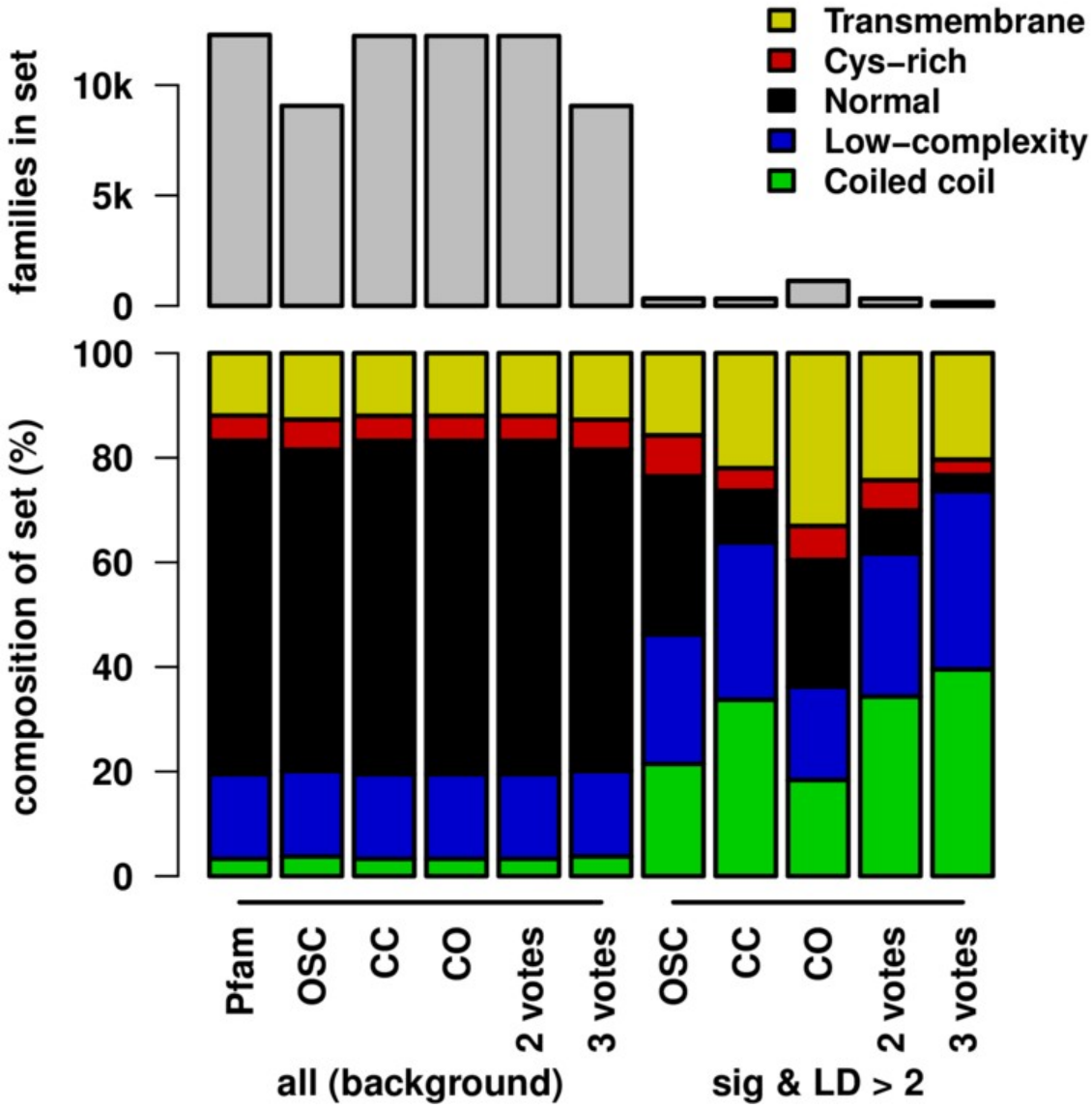


Normal: 7830 (64%)

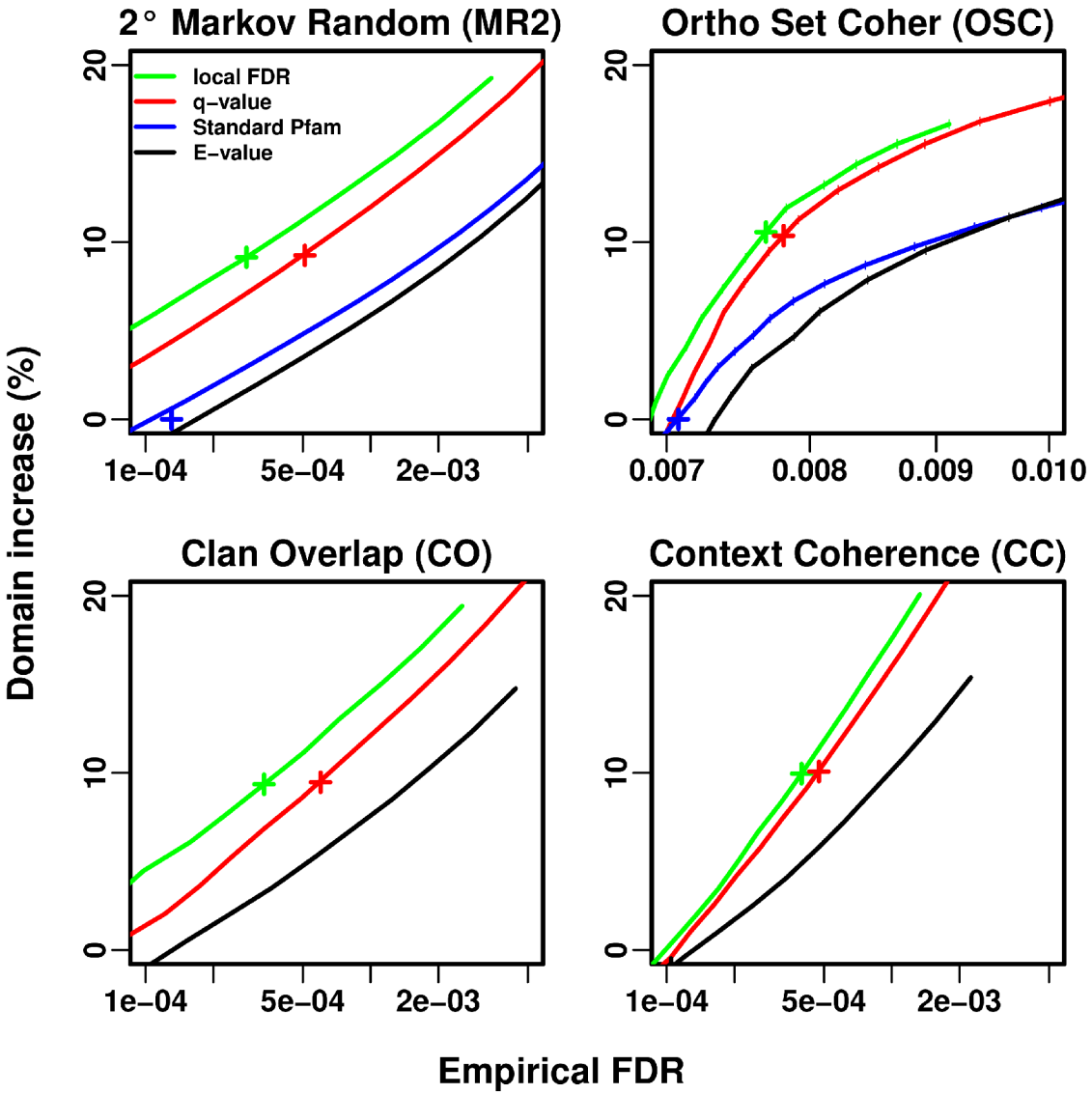
Similar but non-homologous families (different clans)

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

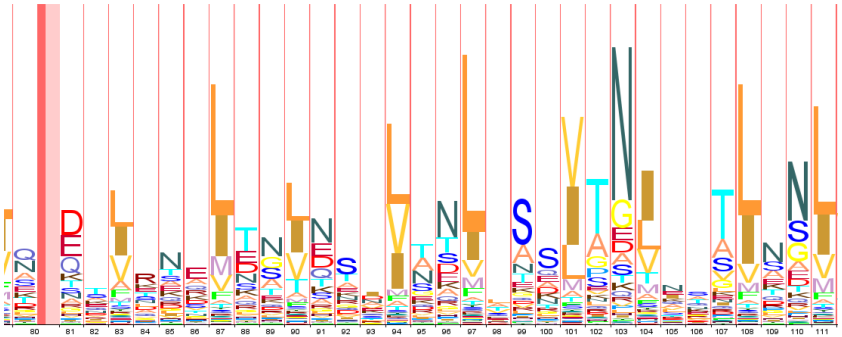
The *q*-value and *local FDR* are better for domains

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

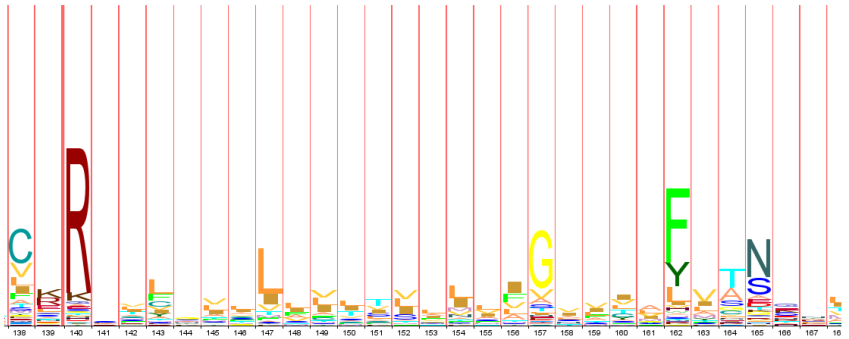
Presented novel empirical null models

- 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).

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
- Application and case studies in *P. falciparum*

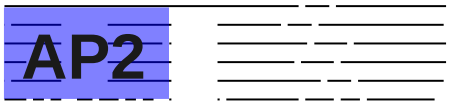
Identification of novel domain families in *Plasmodium falciparum*

15 Apicomplexan proteomes

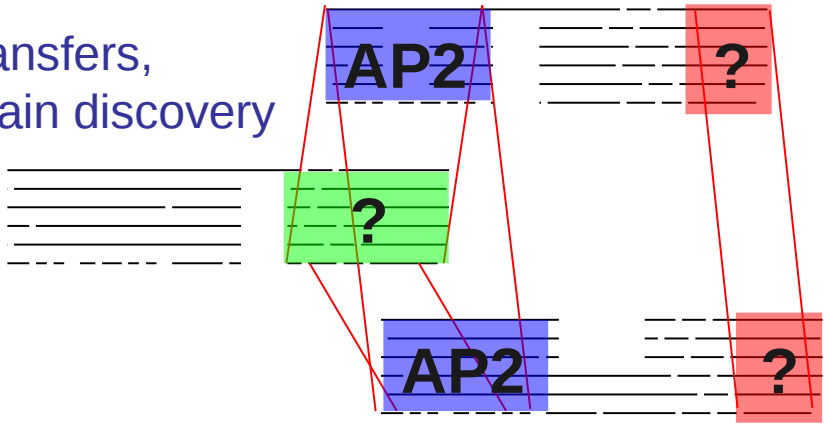
Orthogroup alignments and HMMs

HMM-HMM comparison between orthogroup HMMs

Fancy orthogroup Pfam domain assignments



Domain transfers, novel domain discovery



Pfam domains:
+33% doms, +14% prots
Pfam domain transfer
New non-Pfam domains:
+7.2% doms, +8.7% prots

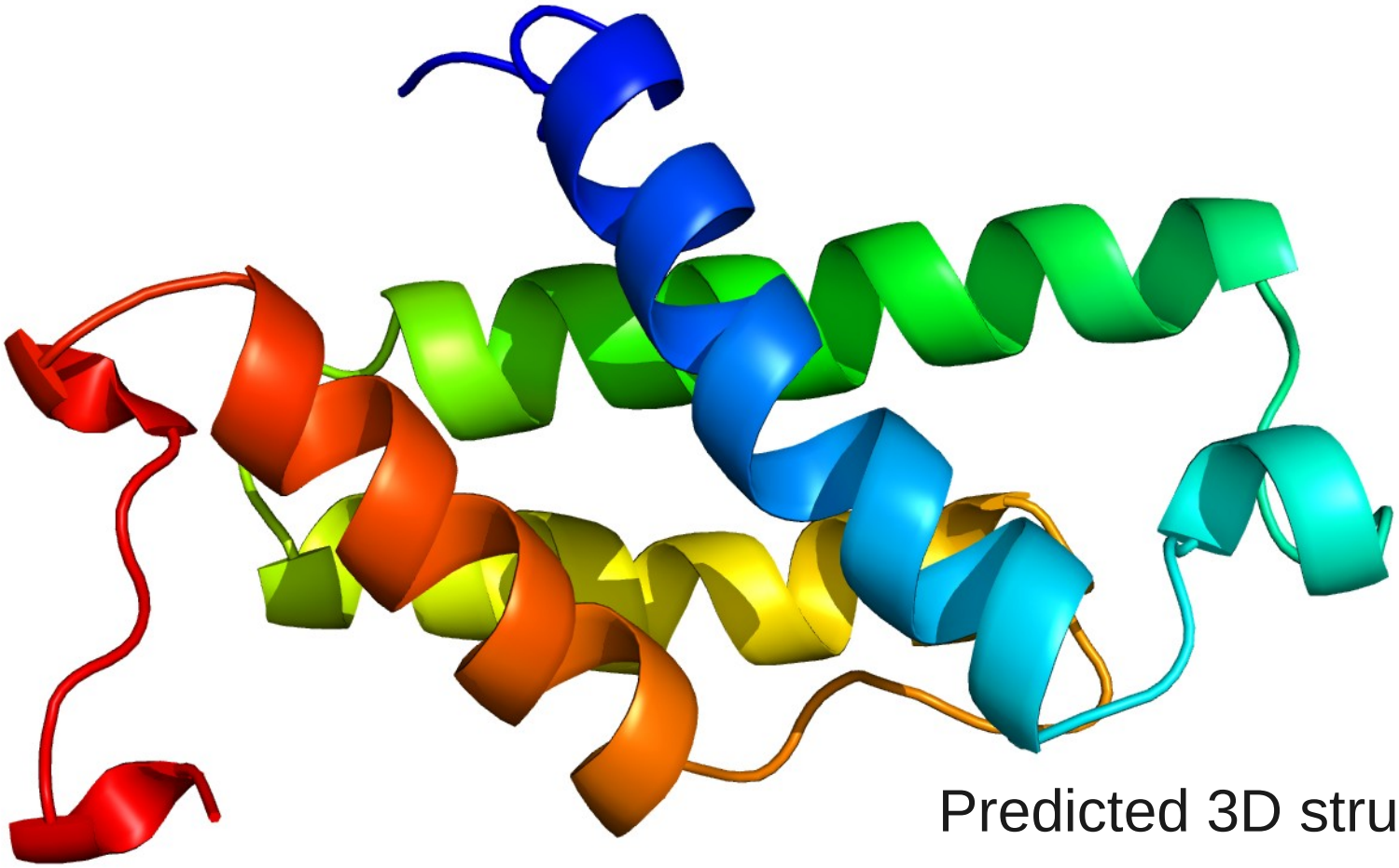
Prediction and experimental validation of novel AP2 domains

Protein	Domain	*q*-value	BILD	Context	Significant motifs
PF13_0026	Dx1	0.0031	*	*	NA
PF13_0114 $	Dx1	0.29		*	NA
	Dx2	0.049	*		NA
MAL7P1.167 $	Dx1	0.31	*		NA
	Dx2	0.4		*	
	Dx1+Dx2	-	-	-	
	Dx3	*q*>0.61	*		NA
PF11_0404	Dx1	0.53		*	NA
	Dx2	0.47		*	NA
	Dx3	*q*>0.61	*		
PF13_0267	Dx1	0.5	*	*	
PF07_0126	Dx1	*q*>0.61	*		NA
PF13_0235	Dx1	*q*>0.61	*		NA

$ = Proteins without strong Pfam predictions

With Ariel Schieler & Sebastian Nasamu

Computational analysis of a novel domain that co-occurs with AP2s



Predicted 3D structure

Follow up experiments by Joana Santos, Mehak Mumtaz & Marcos Lanio

PRINCETON UNIVERSITY

Mona Singh, Computer Science

- Eric Banks
- Tony Capra
- Jesse Farnham
- Dario Ghersi
- Peng Jiang
- Zia Khan
- Elena Nabieva
- Shilpa Nadimpalli
- Anton Persikov
- Yuri Pritykin
- Pawel Przytycki
- Jimin Song
- Josh Wetzel
- Tao Yue
- Elena Zaslavsky

Thesis Committee

- John Storey
- Leonid Krugliak
- Saeed Tavazoie

NSF GRFP
NIH

Manuel Llinás, Molecular Biology

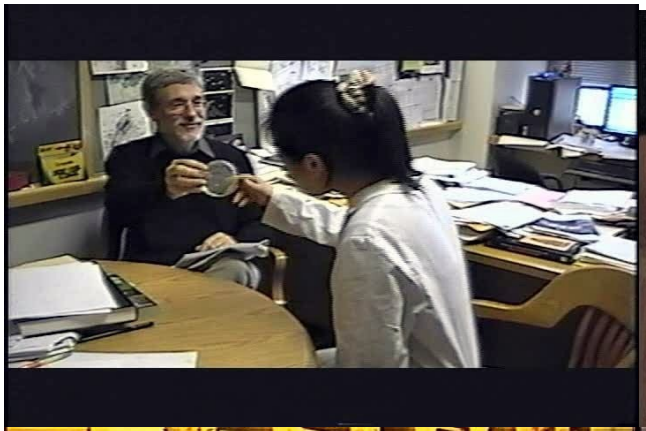
- Lindsey Altenhofen
- Katie Baska
- Tracey Campbell
- Manuela Carrasquilla
- Simon Cobbold
- Valerie Crowley
- Erandi De Silva
- Björn Kafsack
- Elyse Kozlowski
- Lucky Lakshmanan
- Marcos Lanio
- Ilsa Leon
- Ian Lewis
- Yael Marshall
- Mehak Mumtaz (Oxford)
- **Sebastian Nasamu (MIT)**
- Jessica O'Hara
- Kellen Oszewski
- Heather Painter
- Yoanna Pumpalova
- Joana Santos
- Louis Sarry
- **Ariel Schieler**
- Jiang Wang
- Daniel Wilinski
- April Williams
- Irene Ying

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

Singh lab, 2009

Llinás lab, 2011

Molbio skits!









Big Dumb Rock Show


HELICOPTER
SEARCH & RESCUE
GAP
85


ENGLISH
FRENCH
POLISH
STOP


BUY