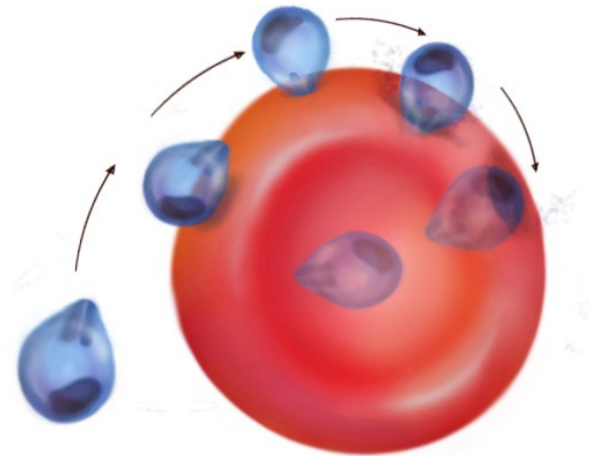
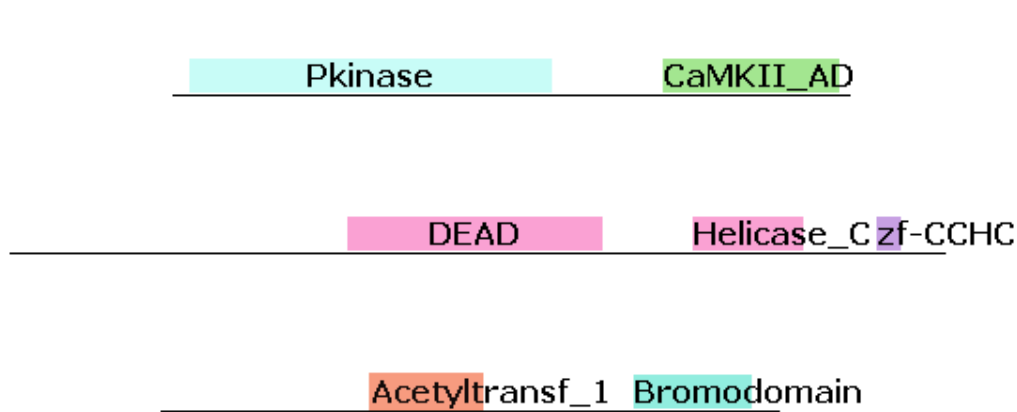
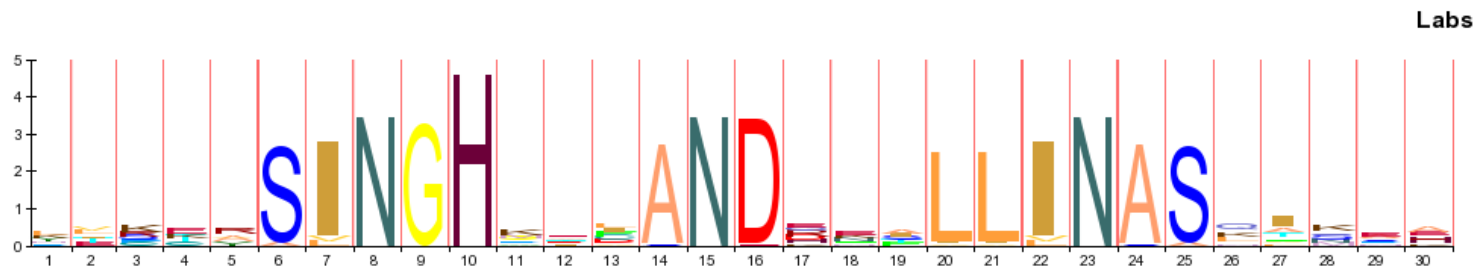


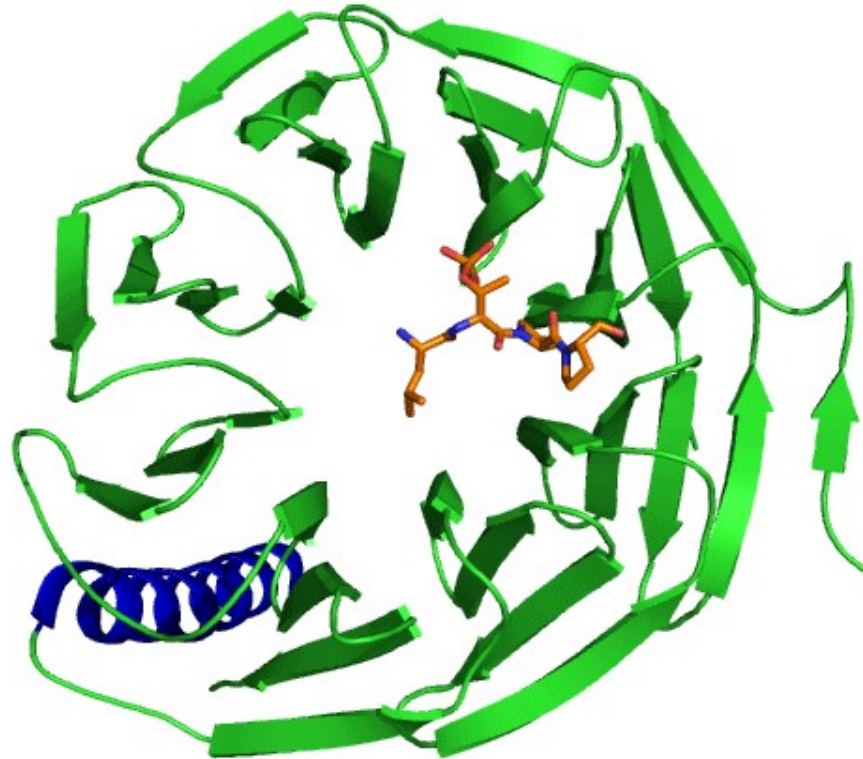
Domain Prediction Using Context in *Plasmodium falciparum*



Alejandro Ochoa
2012-03-01



Protein domains



Domain predictions:

F-box

WD4 WD40 WD40

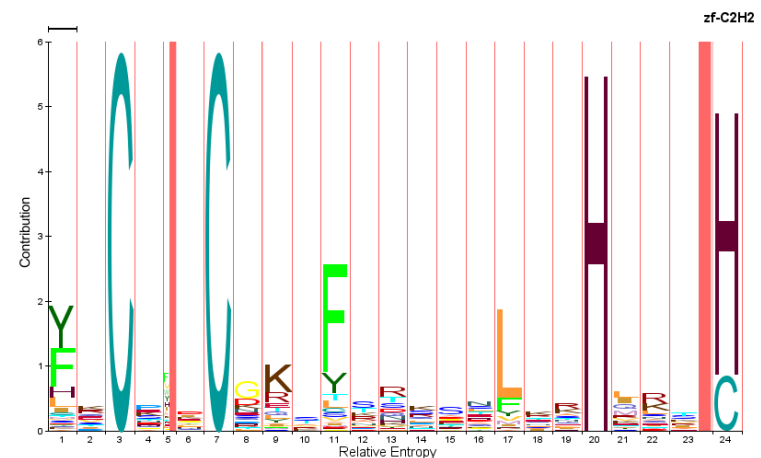
WD40 WD40 WD40

Pfam: a database of protein domain families

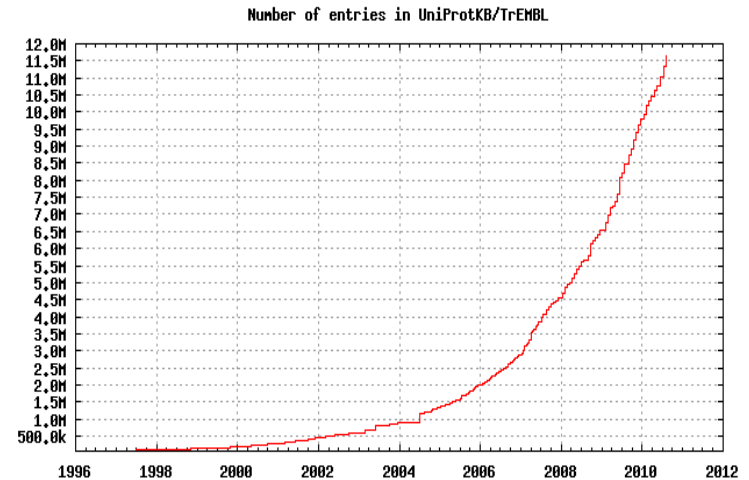
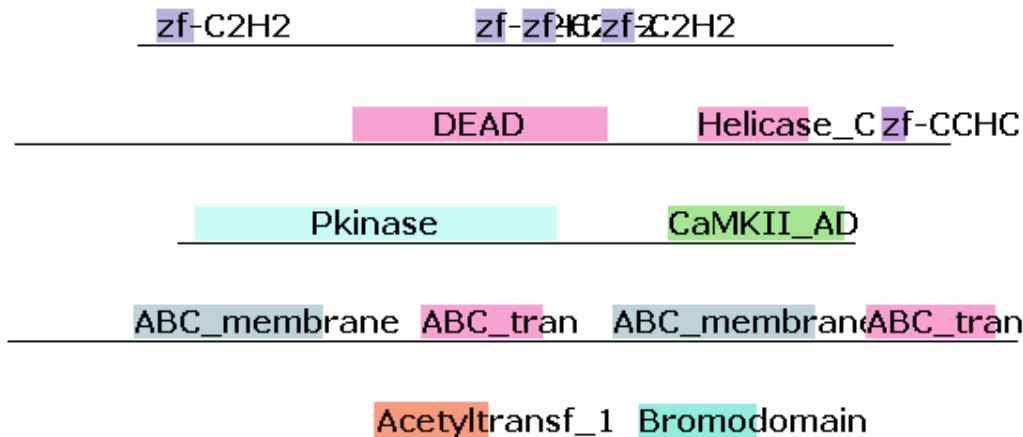
```

SNAI_DROME/362-385      YACQ...VCH...KSF SRM...SLLNKHSSS..NC
SNAI_XENLA/232-255     YQCK...SCS...RTFSRM...SLLHKHEET..GC
SNAI_MOUSE/236-259     YQCQ...ACA...RTFSRM...SLLHKHSES..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
EV11_HUMAN/217-239     IKCK...DCG...QMFSTT...SSLNKHRRF...C
Z02_XENLA/34-59        YSCA...DCG...KHFSEK...MYLQFHQKNPSEC
EV11_HUMAN/21-44       YRCE...DCD...QLFESK...AELADHQKF..PC
ZNF10_HUMAN/517-539    YKCN...QCG...IIFSQM...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
LOLAI_DROME/794-817    YECR...HCG...KKYRWK...STLRRHENV..EC
ZNF17_HUMAN/435-457    YECN...KCG...KFFRYC...FTLMRHQRV...H
ZG32_XENLA/34-56       FVCV...HCG...KGFDRM...YKLSLHLRI...H
TF3A_BUFAM/104-128     YVCYF.ADCG...QQFRKH...NQLKIHOYI...H
ZG46_XENLA/146-168     YVCT...ECG...TSFRVR...PQLRIHLRT...H
MZFI_HUMAN/412-434     FVCG...DCG...QGFVRS...ARLEEHRV...H
ZN239_MOUSE/6-28       YKCD...KCG...KGFTRS...SSLVHHSV...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...DMMTAHVKI..IH
ZNF76_HUMAN/285-309    YTCPE.PHCG...RGFTSA...TMYKNHVRI...H
SDC1_CAEEL/145-168     YMQC...VCL...TLFGHT...YNLFMHURT..SC
SRYC_DROME/358-380     YQCD...ICG...KQFVQK...INLTHHARI...H
SDC1_CAEEL/270-292     YFCH...ICG...TVFIEQ...DMLFKHWRL...H
TRA1_CAEEL/276-300     NKCEY.PCGG...KEYSRL...ENLKTHRRT...H
ESCA_DROME/370-392     CKCN...LCG...KAFSRP...WLLQGHIRT...H
    
```

- 11,912 curated families!
- Profile Hidden Markov Models (HMMs): probabilistic models of sequence families



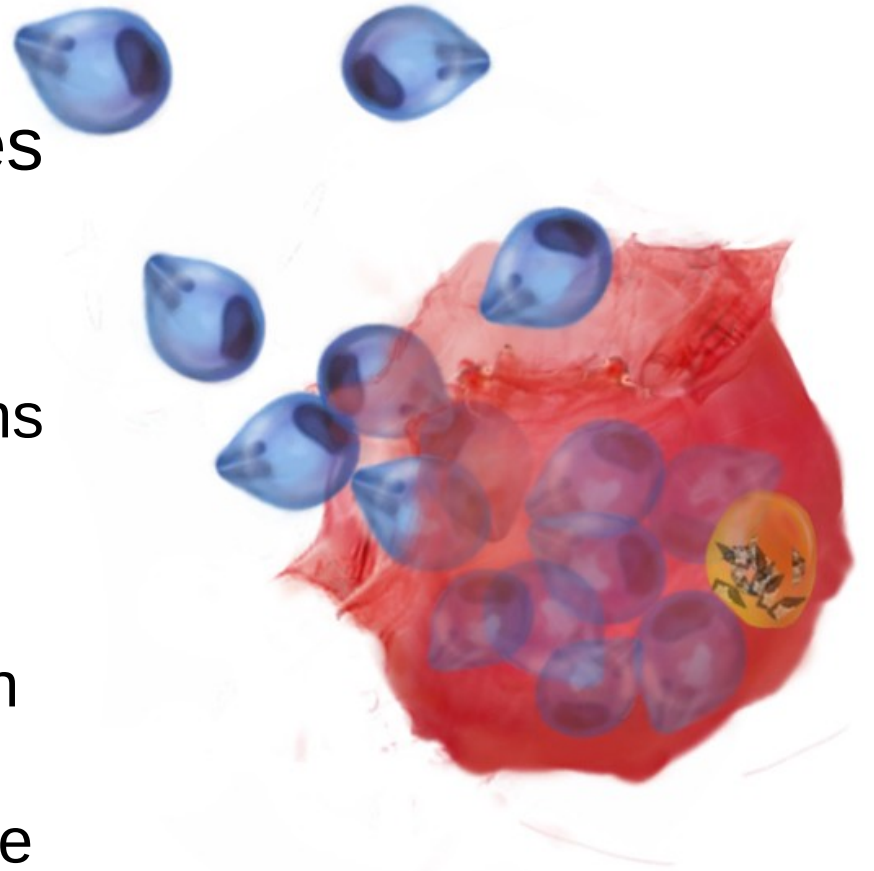
Why predict domains?



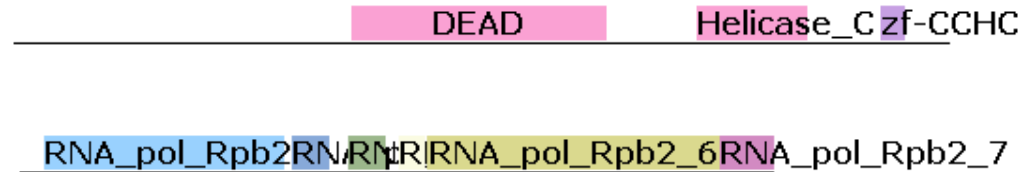
- For new sequences, before experiments start...
- Domains may imply functions
- Experimental alternatives are unfeasible as protein databases grow exponentially

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



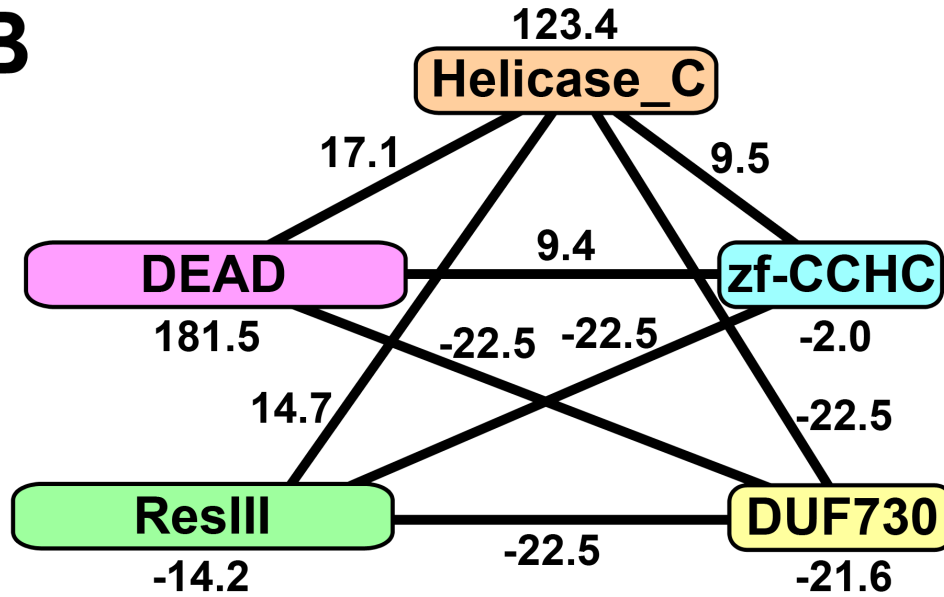
Domain Prediction Using Context: dPUC



- Background
 - Domains co-occur in limited combinations
 - Domains are scored independently of each other
- Idea
 - Score domains in combination
 - Context + Sequence evidence

A

The dPUC
method

B**C**

Standard Pfam:



dPUC Pfam:



General Solution: Integer Linear Programming

$$\text{Max: } \sum_i S_i$$

$$S_i = H_i x_i + \sum_j C_{ij} x_j \quad \forall i \text{ (domain score)}$$

$$x_i, x_j, x_{ij} \in \{0, 1\} \quad \forall i, j,$$

$$0 \leq x_i + x_j - 2 x_{ij} \leq 1 \quad \forall i, j \text{ (} x_{ij} = x_i \& x_j \text{),}$$

$$x_i + x_j \leq 1 \quad \forall i, j \text{ with overlaps,}$$

$$S_i \geq 0 \quad \forall i \text{ (domain thresh)}$$

Speedup: positive elimination

Problem: ILP is too slow with too many domains.

$$S_{i,P}^+ = H_i + \sum_{j \in P} \max \{ 0, C_{ij} \}$$

Eliminate i unless $S_{i,P}^+ \geq 0$, iterate.

Very fast and effective!

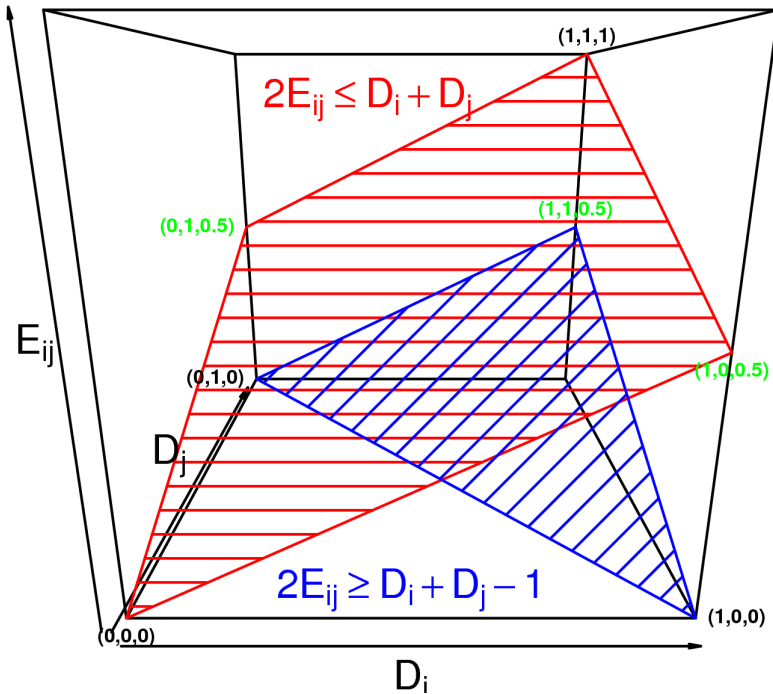
Then solve remaining domains with ILP.

Other speedups (version 2):

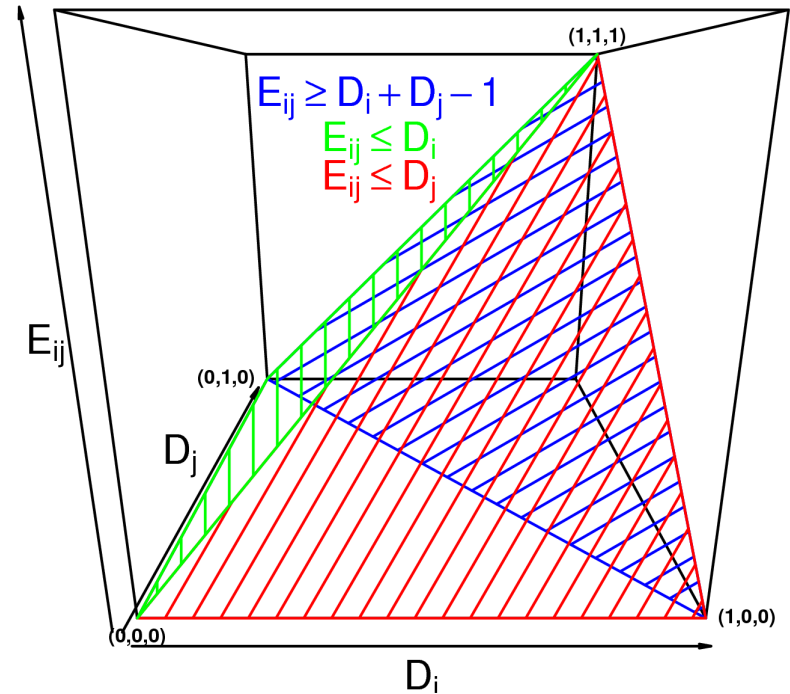
- Trivial cases (no overlaps, all positive context)
- Use C library rather than call executable
- Better constraints

dPUC 2.0: LP constraints

Old $E_{ij} = D_i \& D_j$, $V = 5/12$



New $E_{ij} = D_i \& D_j$, $V = 2/12$

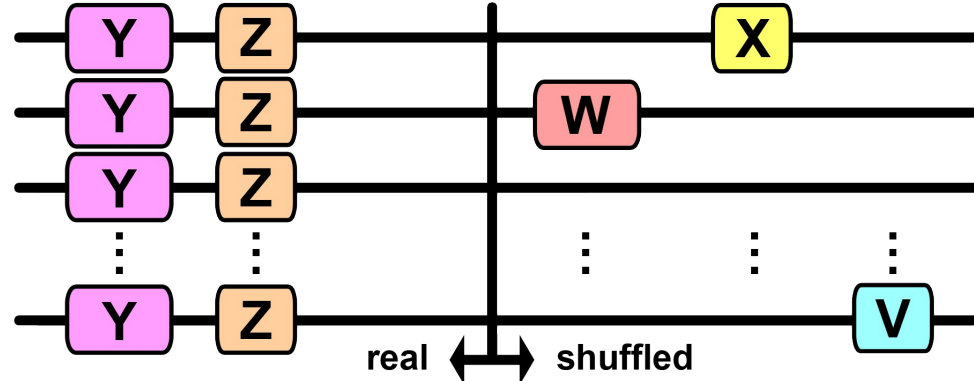


Improved signal to noise

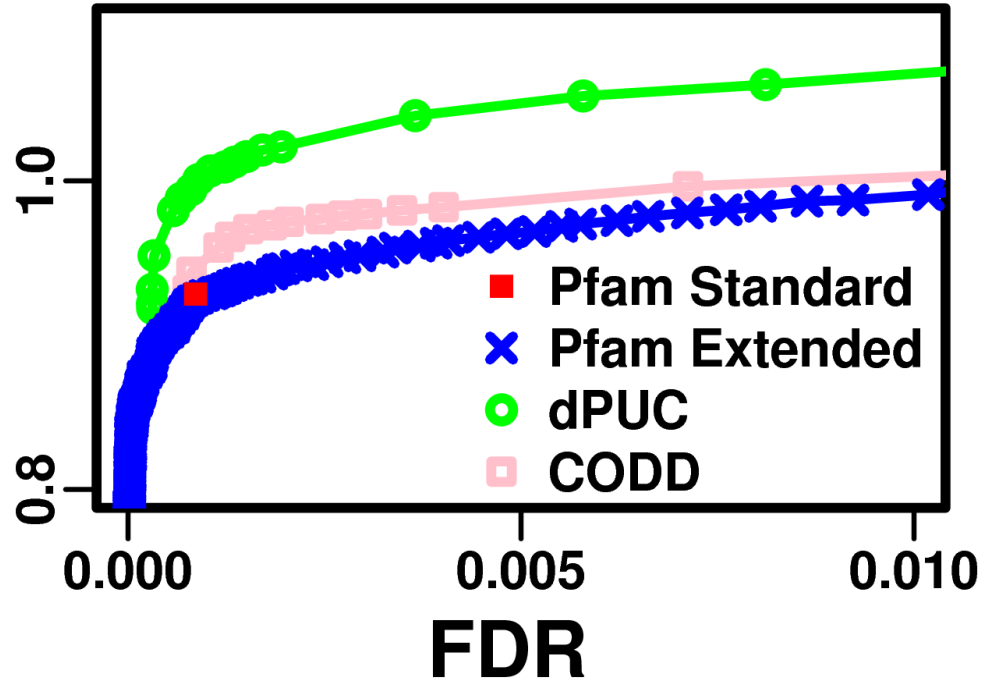
Real protein



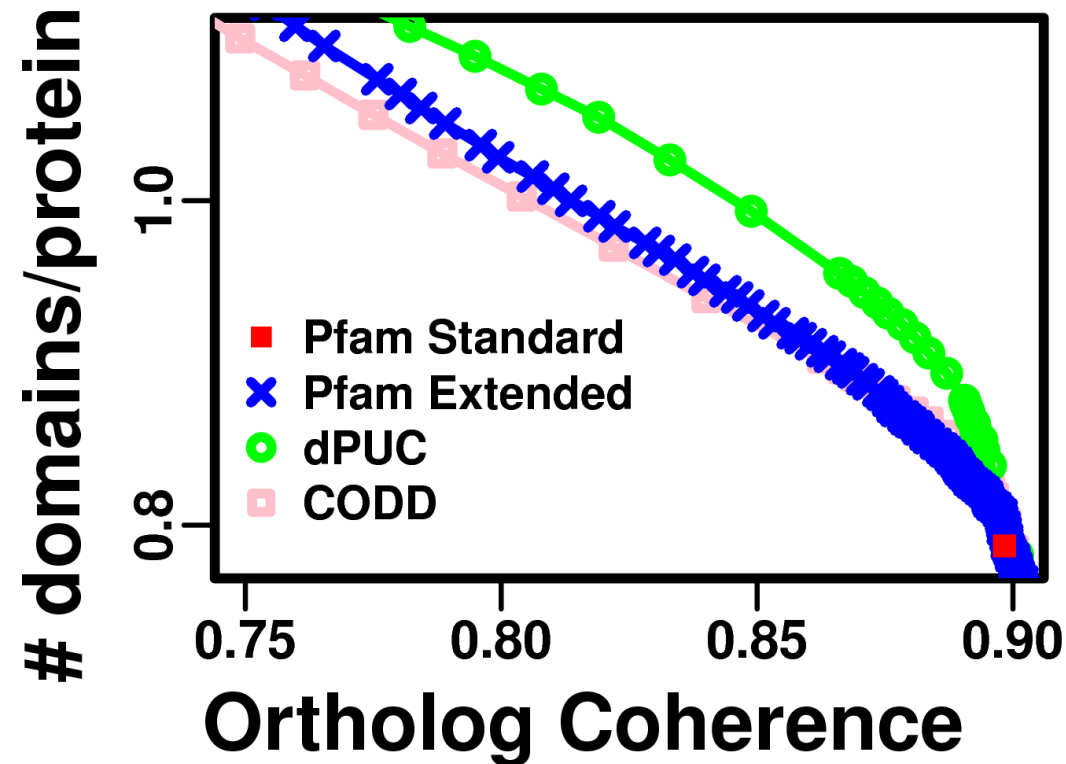
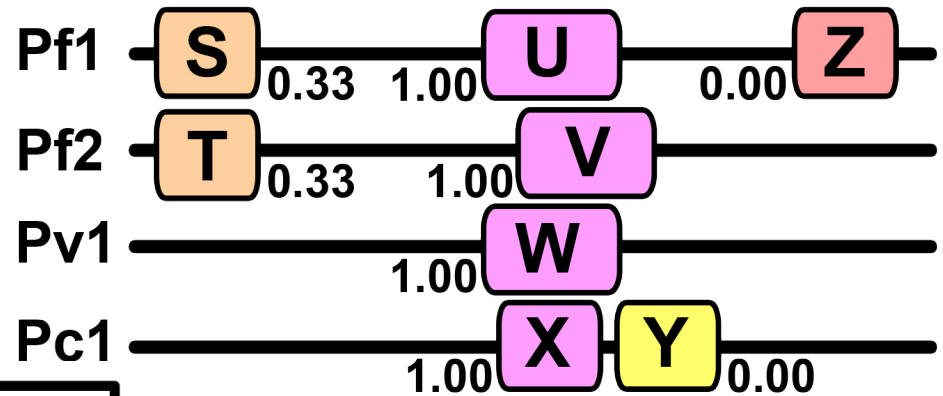
Real protein with shuffled sequences



P. falciparum



Improved ortholog coherence on *Plasmodium* species



New predictions

- Phosphatase -> RNA lariat debranching enzyme
- *P. falciparum*

Standard Pfam
dPUC Pfam

Metallophos
Metallophos **DBR1**







- *S. cerevisiae*

Standard Pfam
dPUC Pfam

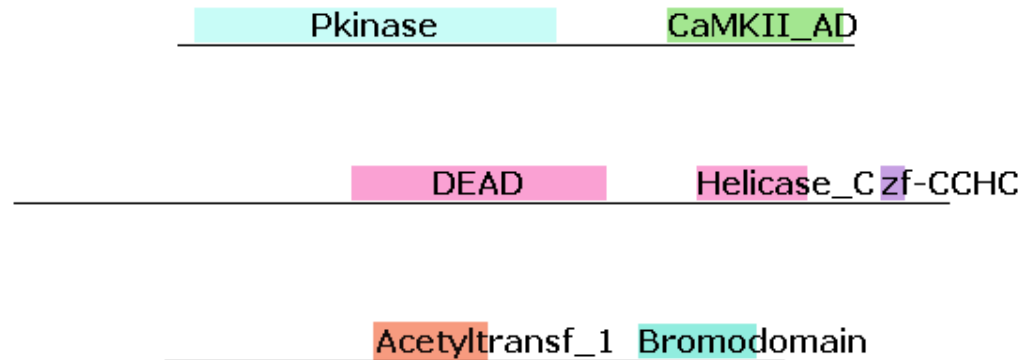
Metallophos**DBR1**
Metallophos**DBR1**

New predictions

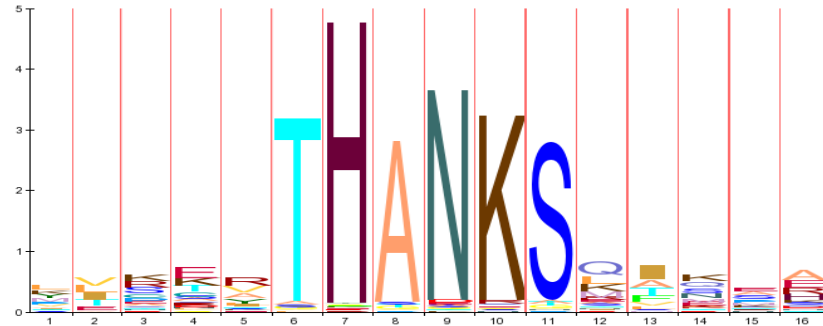
- RNA helicase -> mRNA sequestration

Description	RNA helicase-1
Organism	<i>P. falciparum</i>
Standard Pfam	
dPUC Pfam	
Description	DDX41_DROME ATP-dependent RNA helicase abstrakt
Organism	<i>D. melanogaster</i>
Standard Pfam	
dPUC Pfam	
Description	DDX41_HUMAN Probable ATP-dependent RNA helicase DDX41
Organism	<i>H. sapiens</i>
Standard Pfam	
dPUC Pfam	

Domain context



- Complements sequence evidence
- Improves domain predictions
- Works best on *Plasmodium*

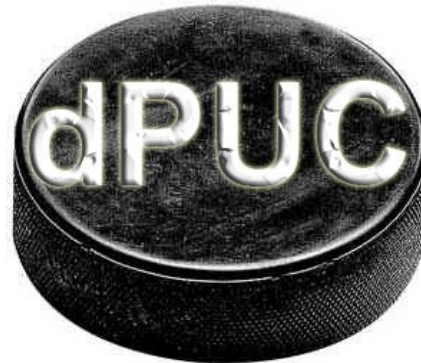


- **Mona Singh & Lab**

- Jesse Farnham
- Dario Ghersi
- Peng Jiang
- Anton Persikov
- Jimin Song
- Josh Wetzel

- **Thesis Committee**

- Leonid Krugliak
- Saeed Tavazoie



dpuc.princeton.edu

- **Manuel Llinás & Lab**

- Lindsey Altenhofen
- Katie Baska
- Simon Cobbold
- Björn Kafsack
- Ian Lewis
- Yael Marshall
- Jessica O'Hara
- Heather Painter
- Joana Santos
- Ariel Schneider
- April Williams

- **NSF GRFP**