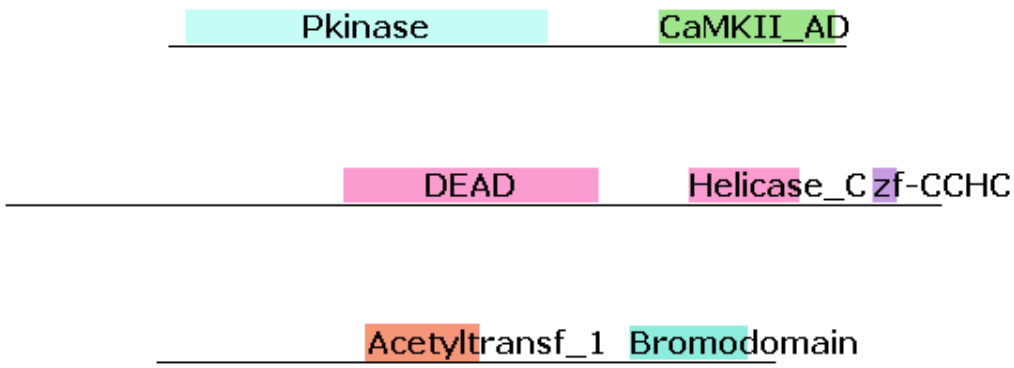
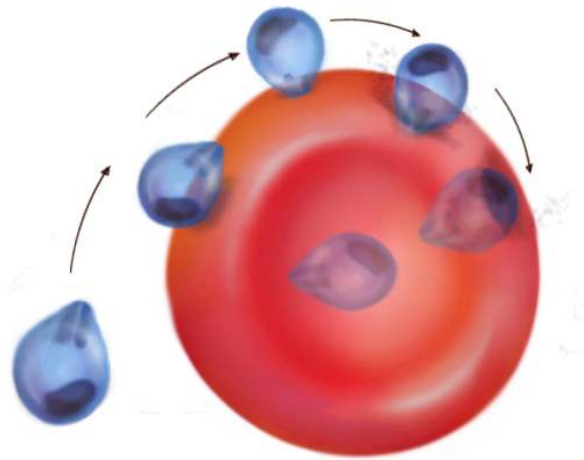
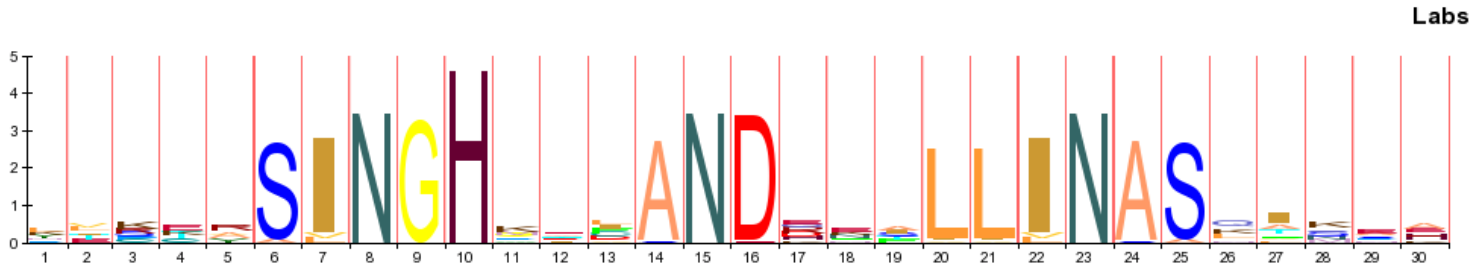
Improving Domain Prediction in *Plasmodium falciparum*



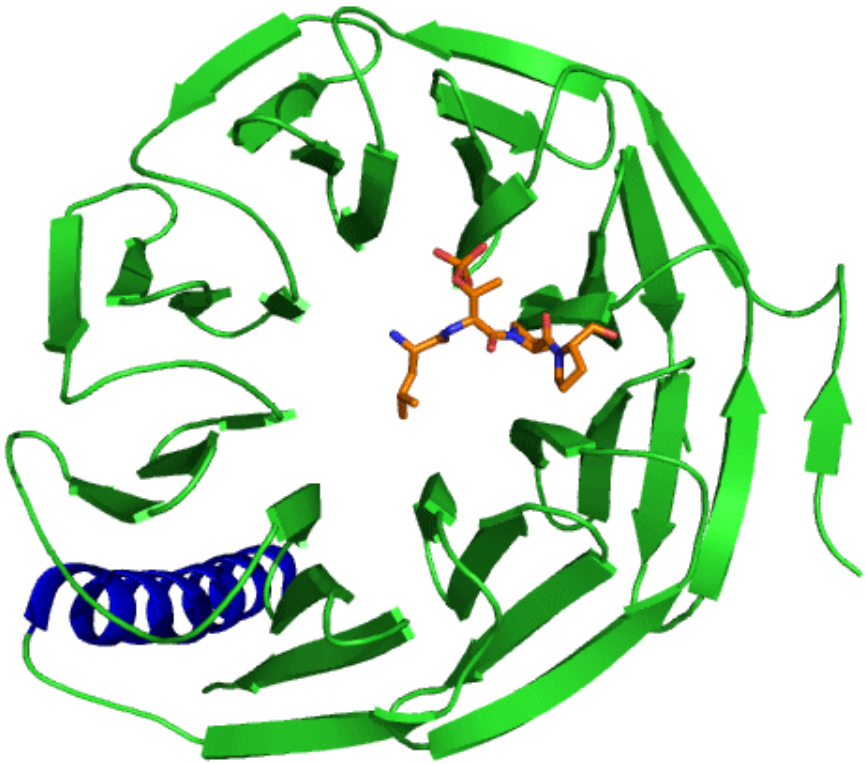


Alejandro Ochoa

2010-10-09



Protein domains



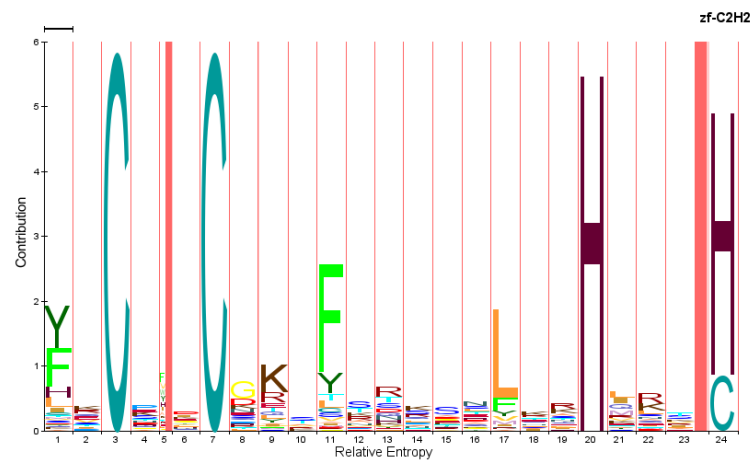
Domain predictions:

Pfam: a database of protein domain families

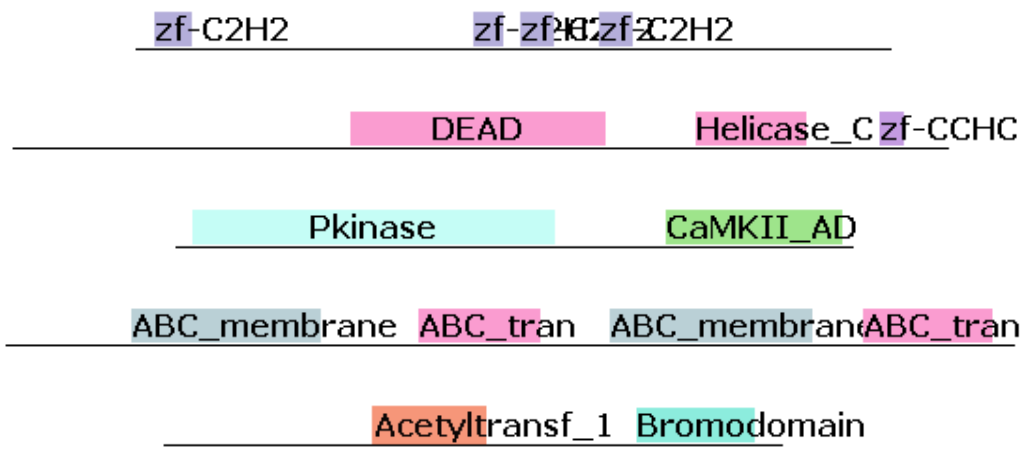
```
SNAI_DROME/362-385      YACQ...VCH...KSFSRM...SLLNKHSSS..NC
SNAI_XENLA/232-255      YQCK...SCS...RTFSRM...SLLHKHEET..GC
SNAI_MOUSE/236-259      YQCQ...ACA...RTFSRM...SLLHKHQES..GC
ESCA_DROME/426-449      YSCT...SCS...KTFSRM...SLLTKHSEG..GC
SUHW_DROAN/221-243      HVCG...KCY...KTFRRL...MSLKKHLEF...C
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
```

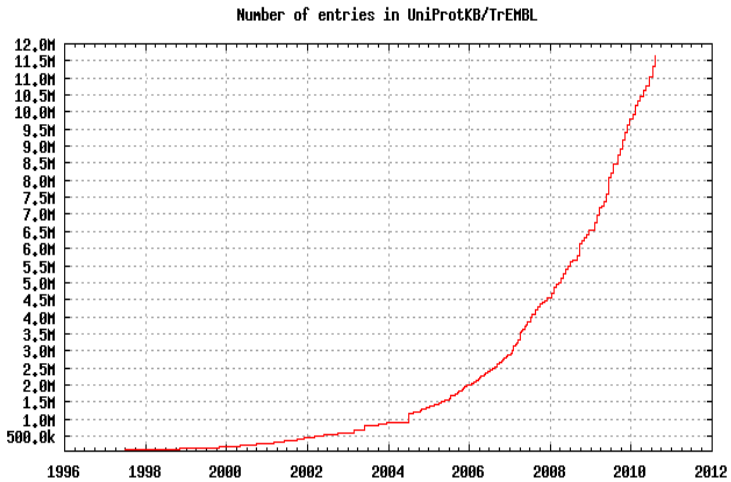
. . . . . .

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



Why predict domains?

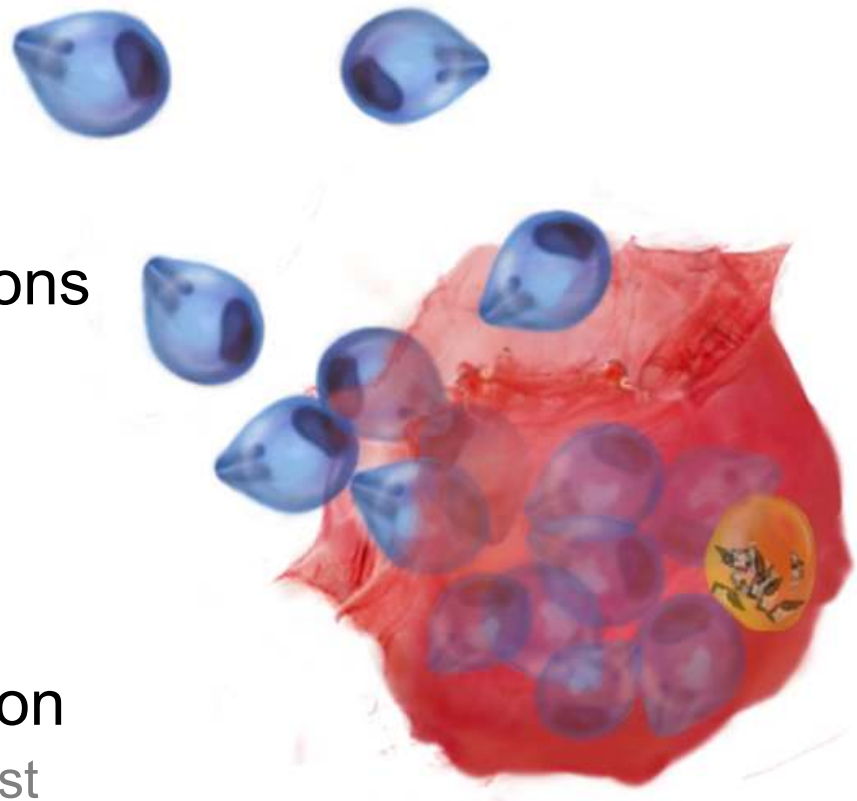




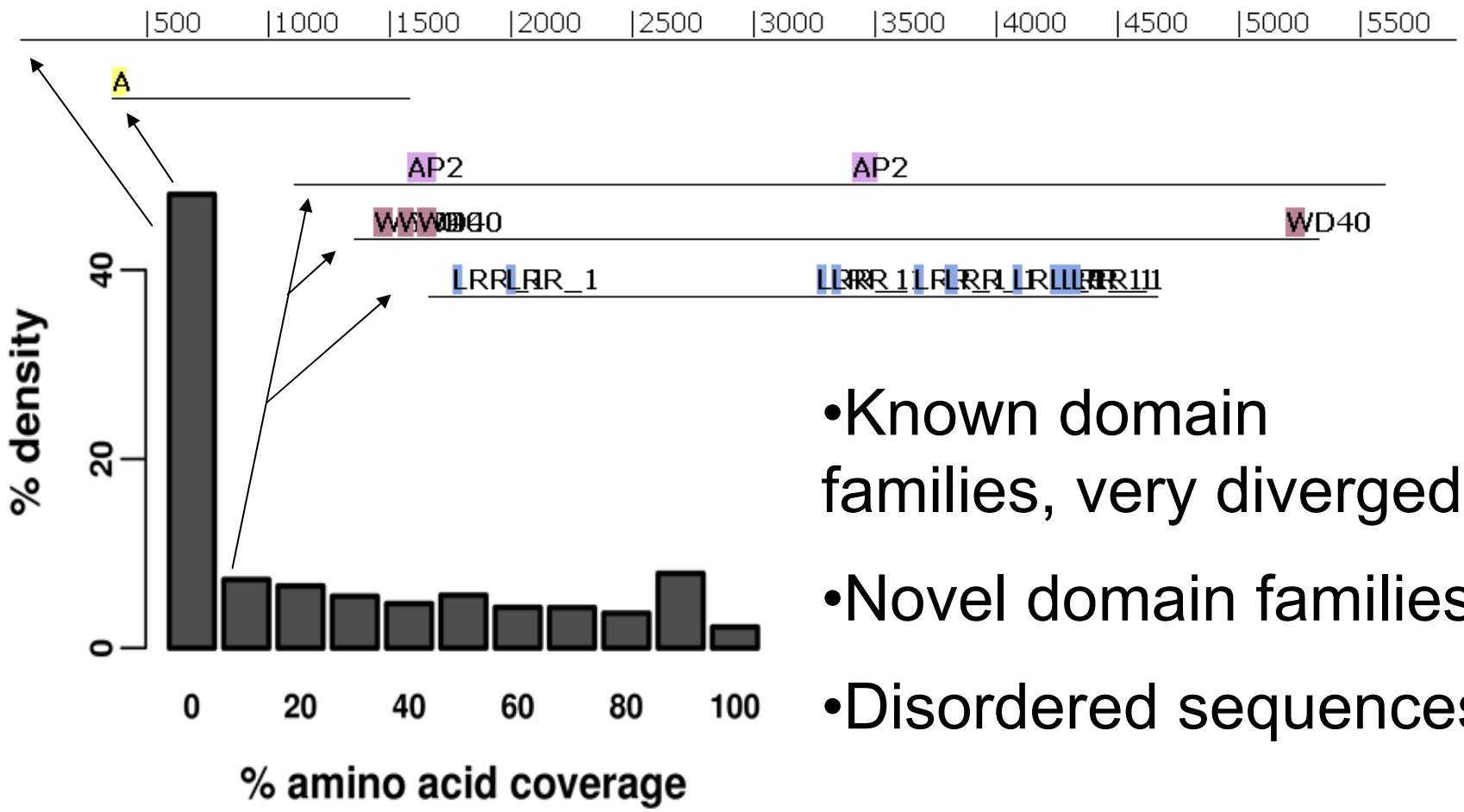
- For new sequences, before experiments start…
- Domains may imply functions
- Experimental alternatives are infeasible as protein databases increase exponentially

Plasmodium falciparum

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

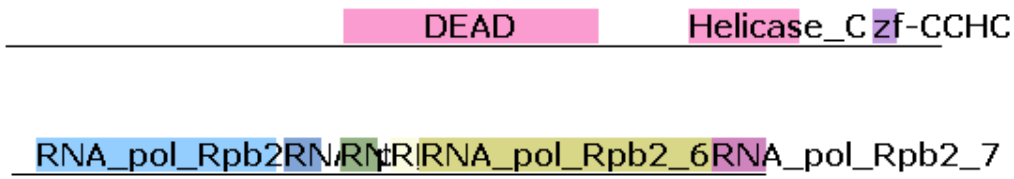


Poor domain coverage of *Plasmodium falciparum*



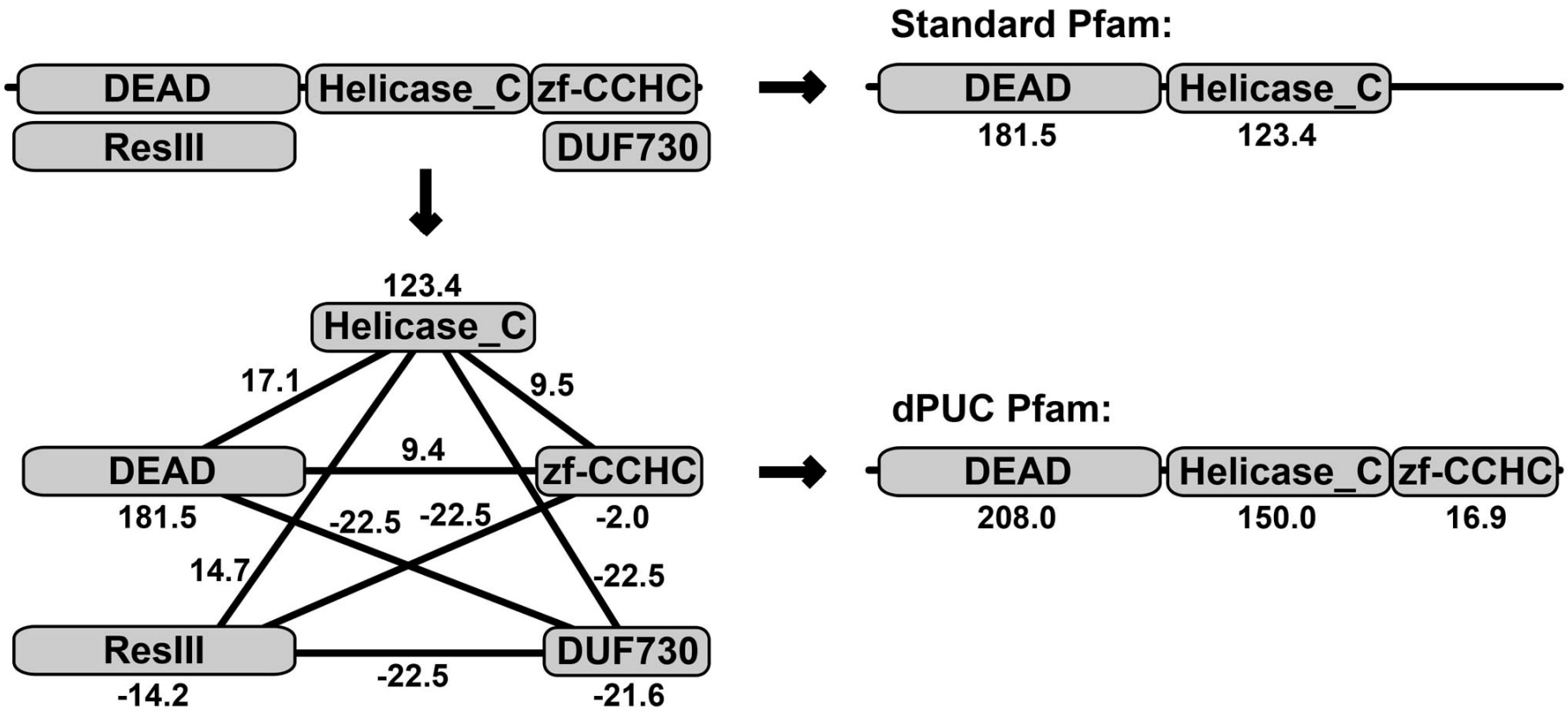
- Known domain families, very diverged
- Novel domain families
- Disordered sequences

Domain Prediction Using Context: dPUC

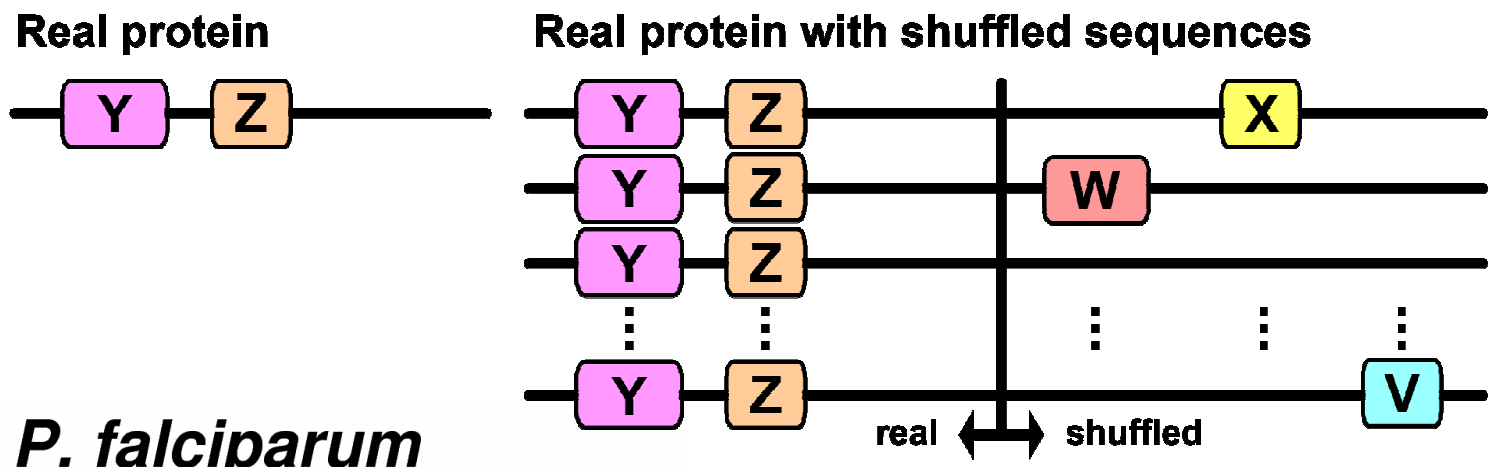


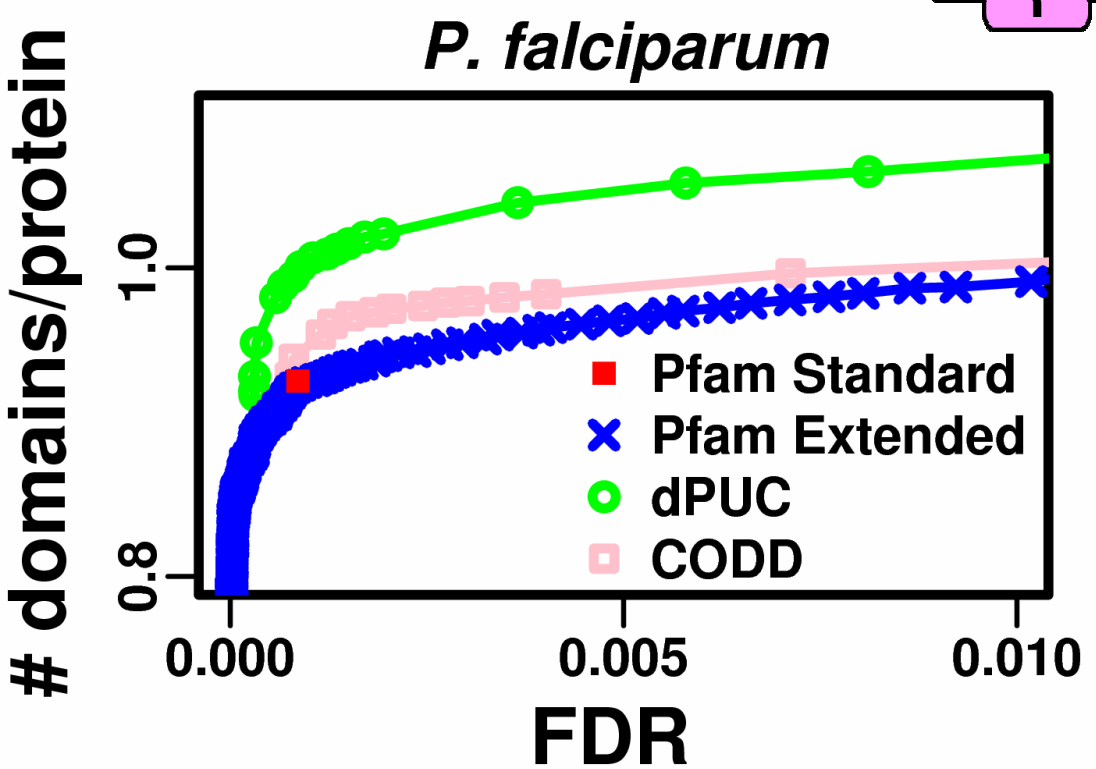
- Background
  - Domains co-occur in limited combinations
  - Domains are scored independently of each other
- Idea
  - Score domains in combination
  - Context + Sequence evidence

The dPUC method

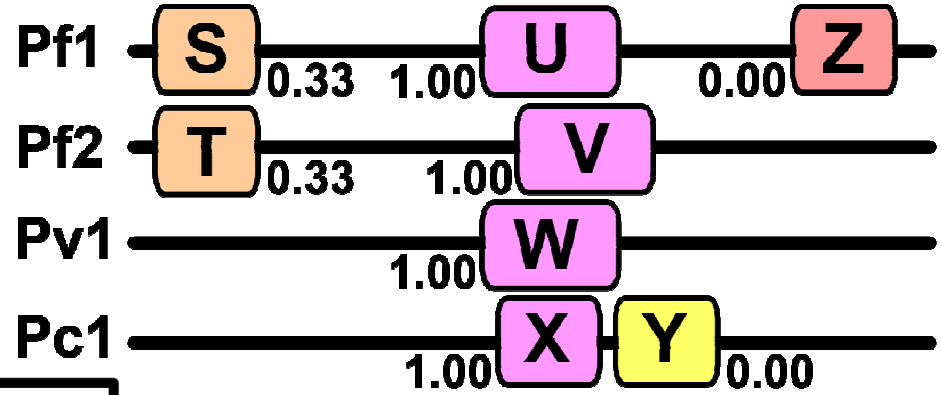


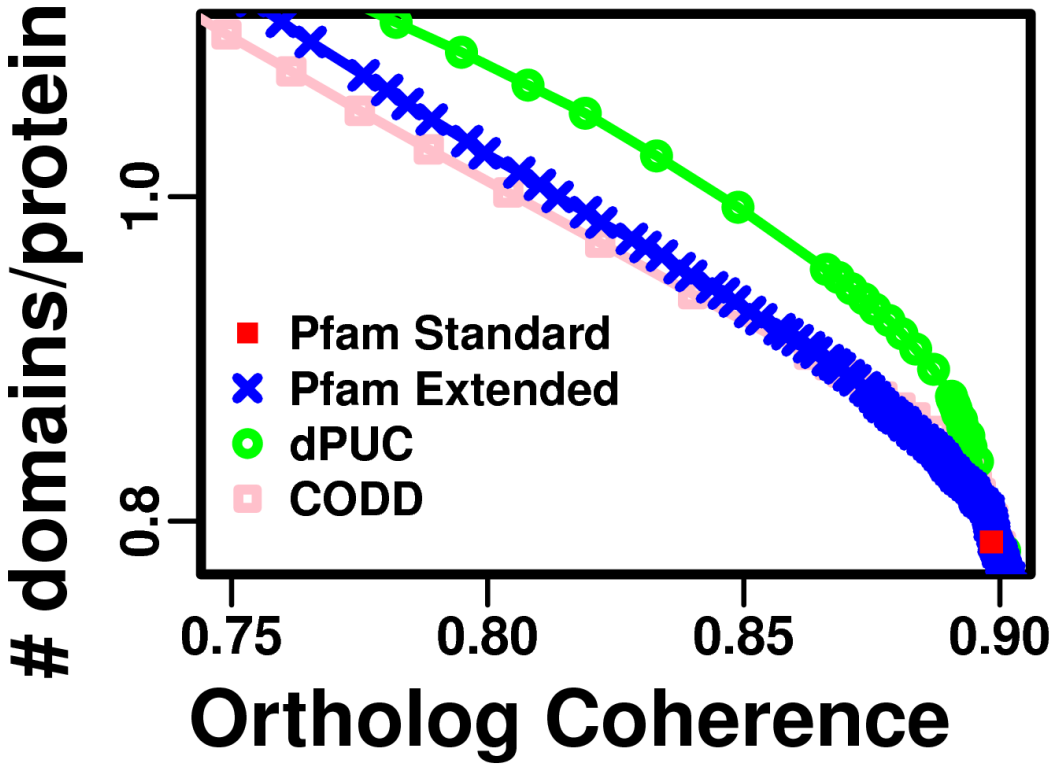
Improved signal to noise



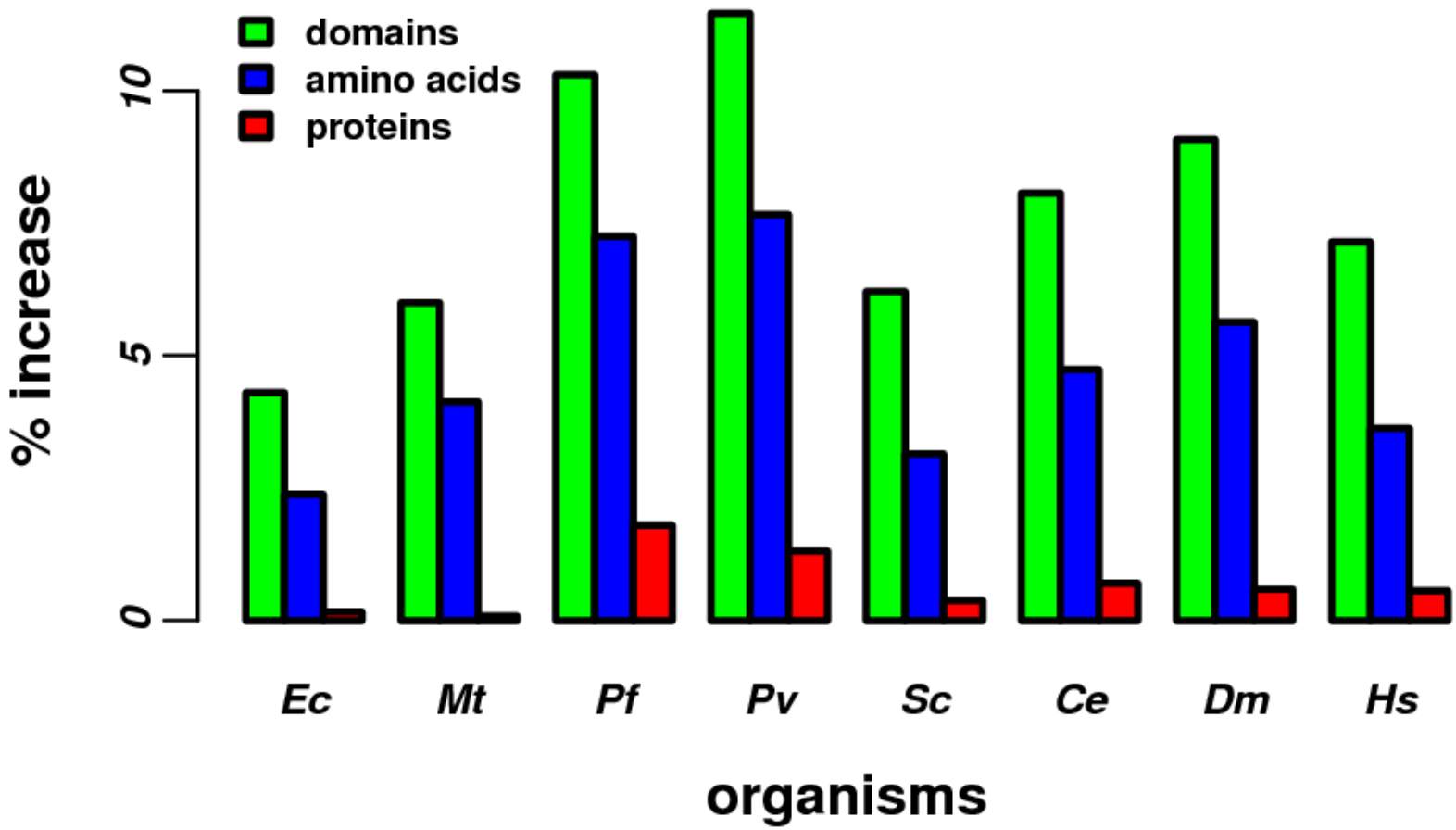


Improved ortholog coherence on *Plasmodium* species



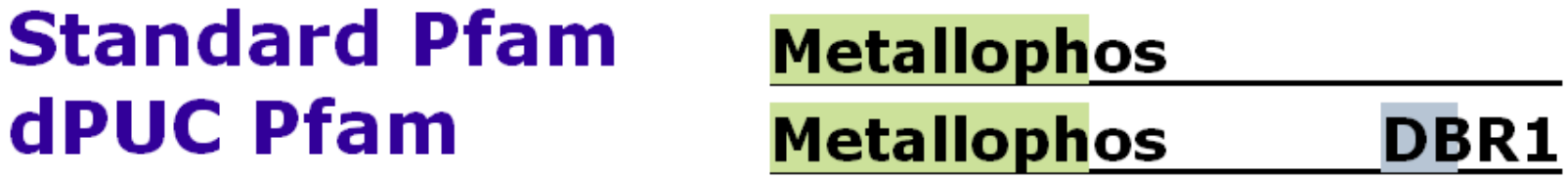


dPUC increases coverage



New predictions

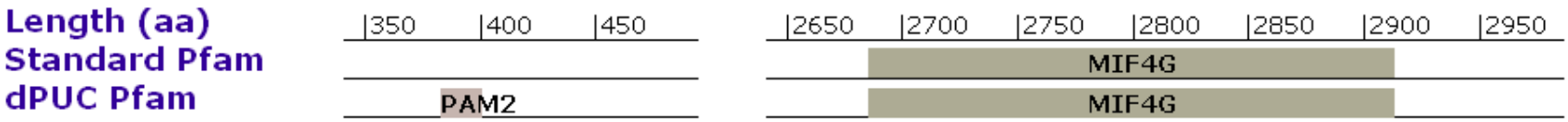
- Phosphatase -> RNA lariat debranching enzyme
- *P. falciparum*



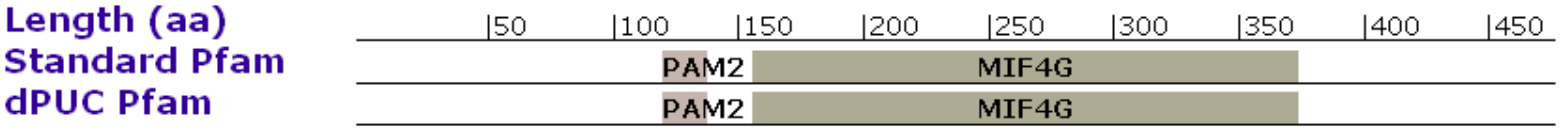
- *S. cerevisiae*

New predictions

- MIF4G domain-containing protein -> Poly-A binding protein-interacting protein 1
- *P. falciparum*



- *H. sapiens*

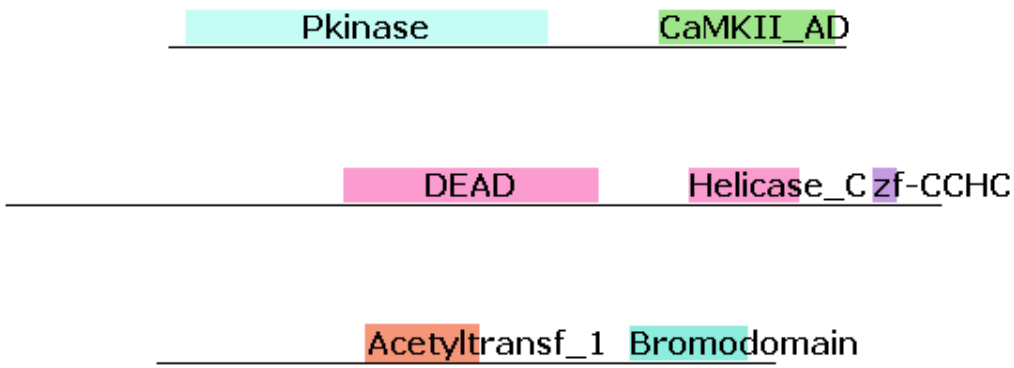


New predictions

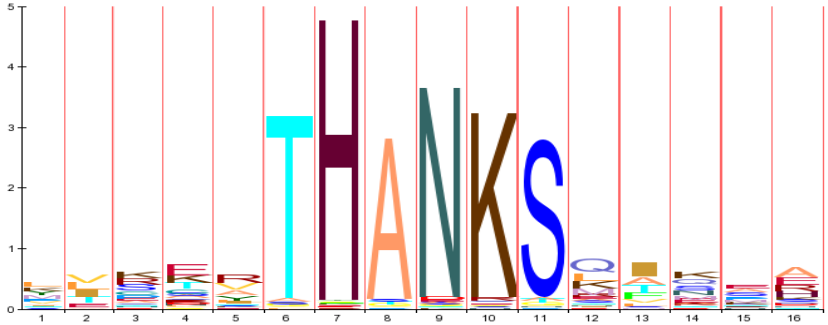
- RNA helicase -> Post-translational mRNA regulation

Description	RNA helicase-1
Organism	*P. falciparum*
Standard Pfam	DEAD Helicase_C
dPUC Pfam	DEAD Helicase_C zf-CCHC
Description	DDX41_DROME ATP-dependent RNA helicase abstrakt
Organism	*D. melanogaster*
Standard Pfam	DEAD Helicase_C
dPUC Pfam	DEAD Helicase_C zf-CCHC
Description	DDX41_HUMAN Probable ATP-dependent RNA helicase DDX41
Organism	*H. sapiens*
Standard Pfam	DEAD Helicase_C zf-CCHC
dPUC Pfam	DEAD Helicase_C zf-CCHC

Domain context



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



- Mona Singh & Lab
  - Jesse Farnham
  - Dario Ghersi
  - Peng Jiang
  - Zia Khan
  - Anton Persikov
  - Jimin Song
  - Tao Yue
- Thesis Committee
  - Leonid Krugliak
  - Saeed Tavazoie

dPUC

dpuc.princeton.edu

- Manuel Llinás & Lab
  - Lindsey Altenhofen
  - Tracey Campbell
  - Erandi De Silva
  - Björn Kafsack
  - Ian Lewis
  - Yael Marshall
  - Jessica O'Hara
  - Kellen Olszewski
  - Heather Painter
  - Yoanna Pumpalova
- NSF GRFP