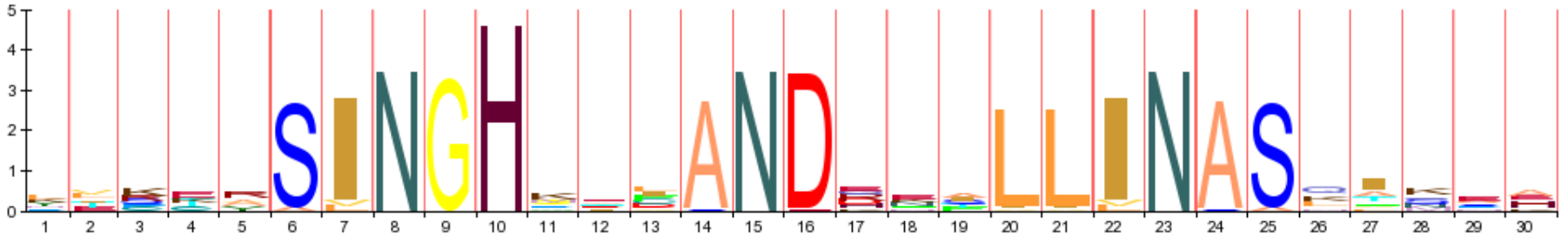
Using domain context for protein domain prediction in *Plasmodium falciparum*

Alejandro Ochoa

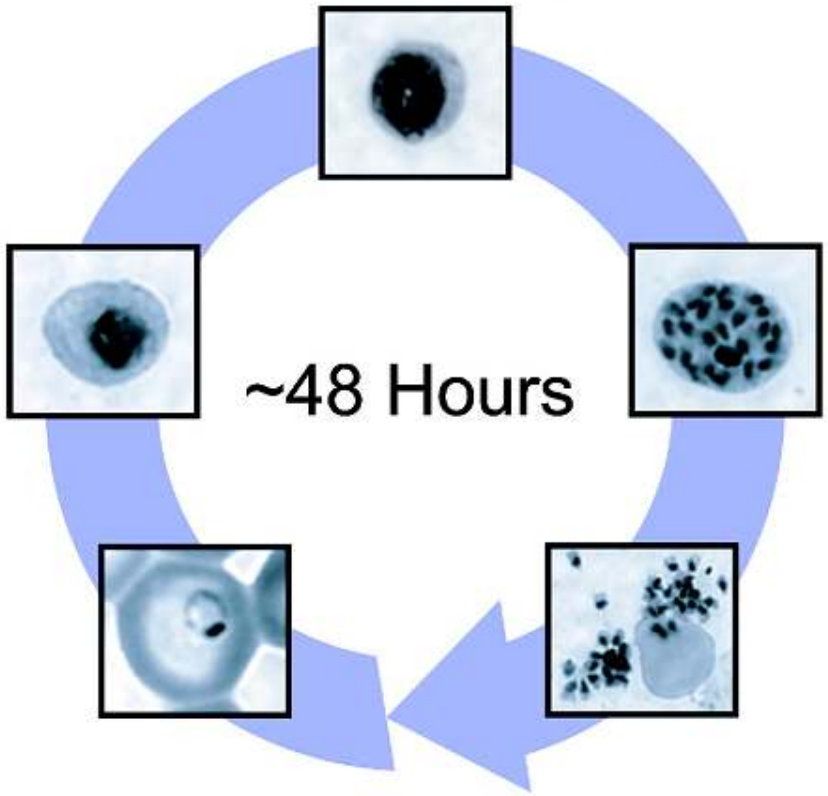
2009-02-05

Labs



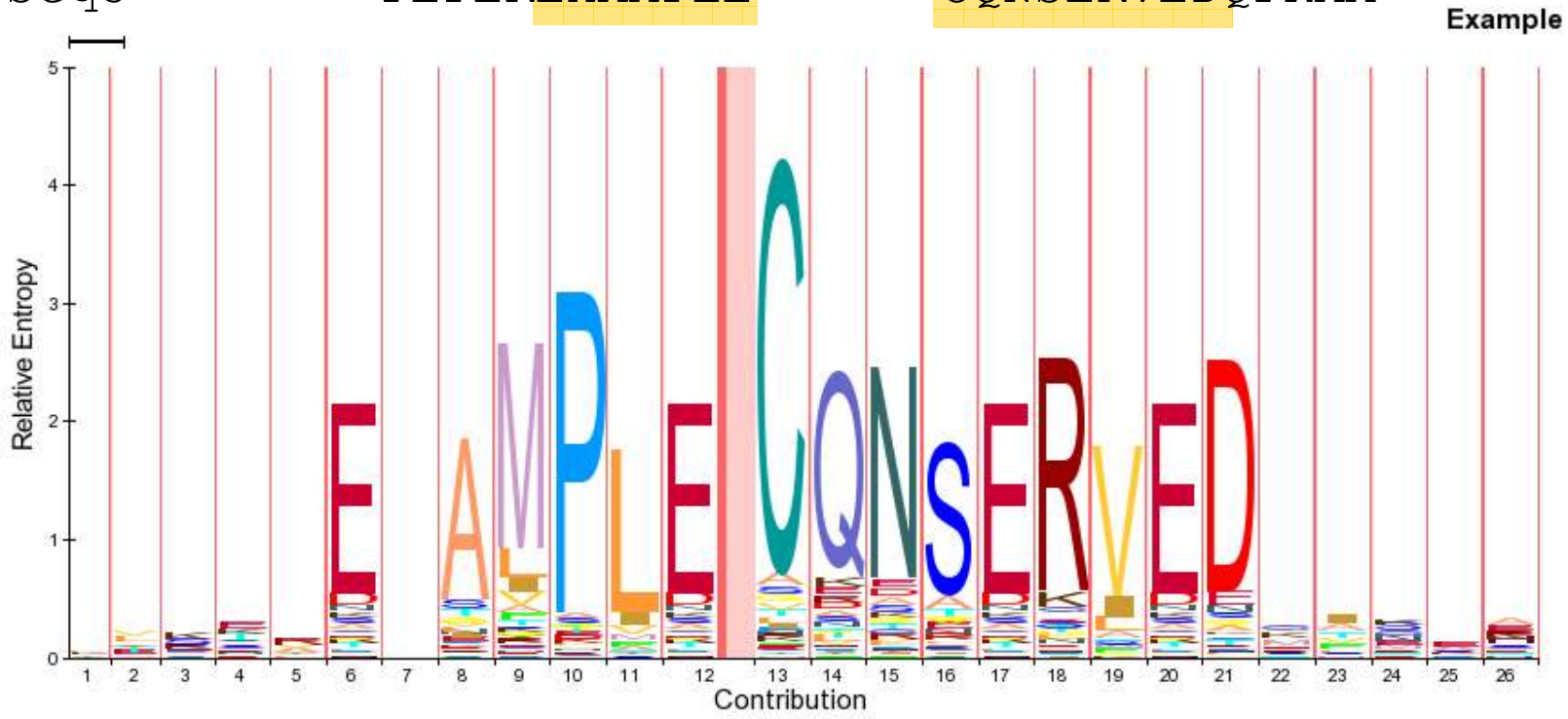
Plasmodium falciparum

- Malaria
  - 515M infections, and
  - 1M deaths per year
  - Drug resistance
- Eukaryote
  - Complex life cycle
- Annotation
  - 5.5K proteins
  - 45% unknown function
    - 20% unknown in yeast
  - 88% of annotations are bioinformatical
  - Some plant-like genes (AP2)



Hidden Markov Models

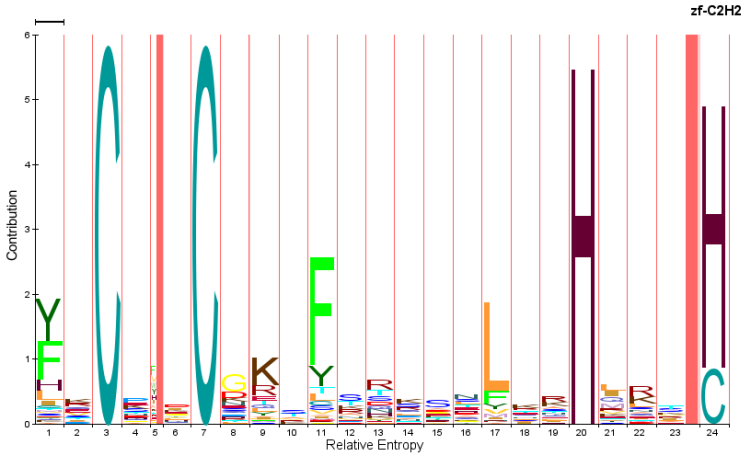
```
Seq1        KLSTAEXAMPLE------CQNSERVEDMINER
Seq2        YTRCVEXAMPLE------CQNSERVEDLAMAH
Seq3        MIKEREXAMPLEINSERTCQNSERVEDKISSA
Seq4        LVCKYEXAMPLE------CQNSERVEDQTQPE
Seq5        PEPEREXAMPLE------CQNSERVEDQFKMA
```



Hidden Markov Models

```
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
```

- C2H2 zinc finger
- Pfam alignment, 192 sequences, 18 organisms
- Expert curated
- HMM recognizes 52,611 sequences in 678 organisms



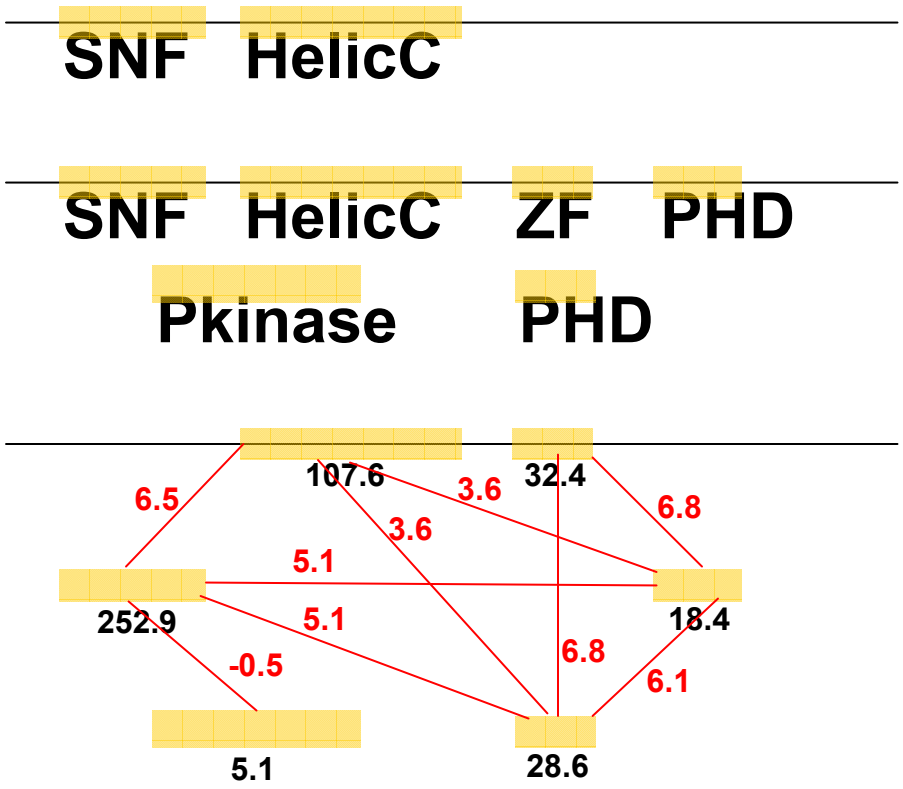
HMM databases

- Pfam
  - 10,340 HMMs, very well curated
  - Describes protein families, domains, repeats, and motifs
- Smart
  - 810 HMMs, sometimes better models than their Pfam counterparts
  - Concentrates on extracellular modules and signaling domains
- Superfamily
  - 1,776 HMMs, less curated
  - Describes protein domain folds, superfamily level

Domain context

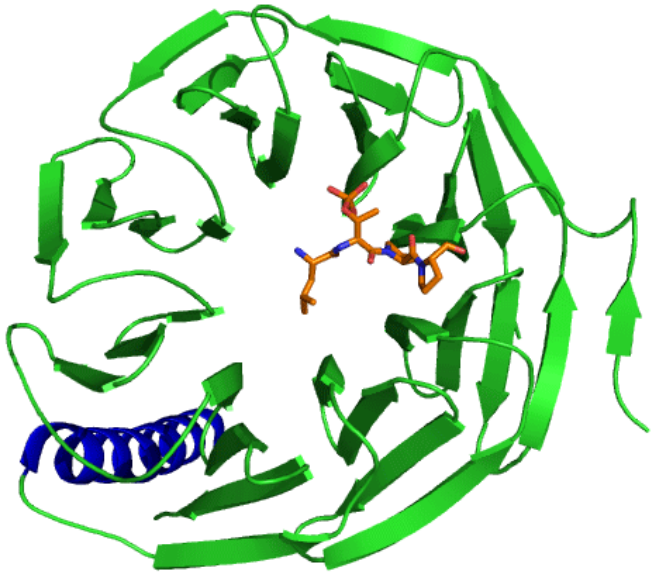
- Background
  - 80% of eukaryotic proteins have many domains
  - Domains are scored independently of each other
- Idea
  - Score domains by taking into account the other domain predictions in the same protein
  - Create fast and practical method that can be applied to genomes and on web
- Precedent
  - **Enhanced protein domain discovery by using language modeling techniques from speech recognition.** L Coin, A Bateman, R Durbin. PNAS 2003; 100(8): 4516–4520

Domain context analysis framework



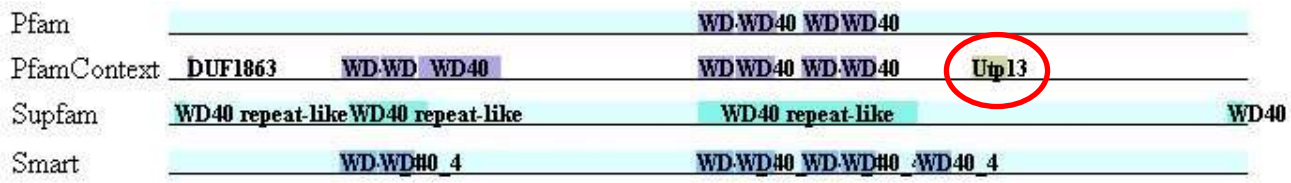
- Pfam without context predicts a limited amount of domains
- Pfam hits w/ E < 1000
- Scores on nodes and edges
- Get non-overlapping domains that maximize score

Context predictions: WD40 + Utp13



PDB: 1NEX

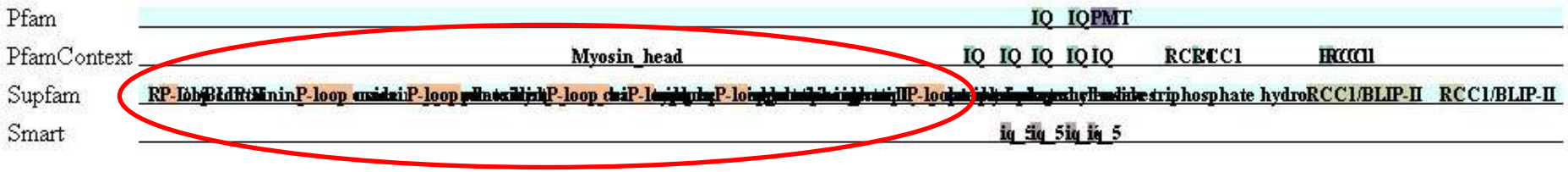
- PVX_094875
  - hypothetical protein, conserved
  - WD40 repeats do not imply function
  - Utp13 implies specific function:
    - u3 small nucleolar rna-associated protein 13
  - Predicted in close orthologs!
  - Same domain order as yeast's!
  - Orthology confirmed with OrthoMCL



New predictions at a glance (incomplete)

- U3 complex (all mostly WD40)
  - PF10_0128: Utp13
  - PF08_0130: Utp1/Pwp2
  - PF14_0456: Utp12/Dip2
- Putative DNA binding/chromatin remodeling
  - PF11_0404: 3+**3** AP2
  - PF14_0079: **3** AT_hook, 1 AP2
  - PF10_0079: 2 AT_hook, 4+**1** PHD
  - PF14_0409: **2** AT_hook
  - PFL0465c: 7+**4** zf-C2H2, **1** AT_hook
  - PFF1440w: **1** HMG, 3 PHD, 1 Bromo, 1 SET
  - PFF1185w: **3** HMG, 5+**1** PHD, 1 SNF2_N, 1 Helicase_C
- More controversial DNA-binding (low complexity repeats)
  - PFL1930w: **13** bZIP, **1** RRM, **1** zf-C2H2
  - PFI0565w: **3** NUMOD3
  - PFI1595c: **3** NUMOD3
  - PFB0915w: TM, **12** HHH
  - MAL13P1.137: **7** Transposase_8
- Miscelaneous
  - PFE1390w (RNA helicase-1): 1 DEAD, 1 Helicase_C, **1** zf-CCHC
  - PFL0975w: **1** Myosin_head, 2+**3** IQ, 2+**2** RCC1
  - PFE0540w: **1** LisH, 7 WD40

PVX_123655: hypothetical protein, conserved



- Chro.70483 (Cryptosporidium hominis)

- TP01_1107 (Theileria parva)

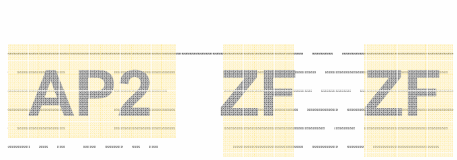
- GSPATT00001002001 (Paramecium tetraurelia)

- MYO12 (Tetrahymena thermophila)

In the future

- Finish math formulation of problem
- Finish Pfam predictions
- Repeat analysis with Superfamily
- Validate predictions

- Singh Lab
  - Mona Singh
  - Eric Banks
  - Tony Capra
  - Peng Jiang
  - Zia Khan
  - Anton Persikov
  - Jimin Song
  - Tao Yue



- NSF GRFP

- Llinás Lab
  - Manuel Llinás
  - Tracey Campbell
  - Erandi De Silva
  - Elyse Kozlowski
  - Viswanathan Lakshmanan
  - Yael Marshall
  - Jessica O'Hara
  - Kellen Olszewski
  - Louis Sarry
  - Jiang Wang
  - Irene Ying
  - Erin Bush
  - Matt Foglia