BCB 5200 Introduction to Bioinformatics

Hidden Markov Models (HMMs): Probabilistic Models

Bioinformatics and Computational Biology

Saint Louis University

Outline

- Markov models and HMMs
- Protein profile HMMs
- HMM tools
- Available resources for Profile HMMs

Markov Model

- Markov Model is part of the theory of probabilities, a stochastic model to model randomly changing systems.

- Set of states: $\{s_1, s_2, \ldots, s_N\}$

- Process moves from one state to another generating a sequence of states :

$$s_{i1}, s_{i2}, \ldots, s_{ik}, \ldots$$

- Markov chain property: probability of each subsequent state depends only on what was the previous state:

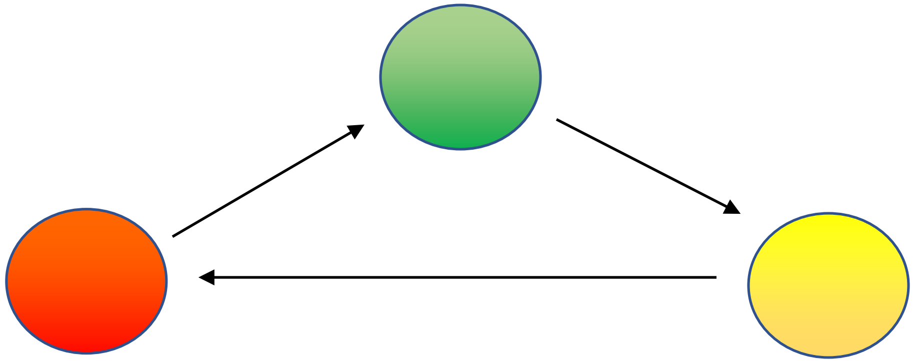
$$P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1}) = P(s_{ik} \mid s_{ik-1})$$

- To define Markov model, the following probabilities have to be specified: transition probabilities $\pi_i = P(s_i)$ and initial probabilities $a_{ij} = P(s_i \mid s_j)$

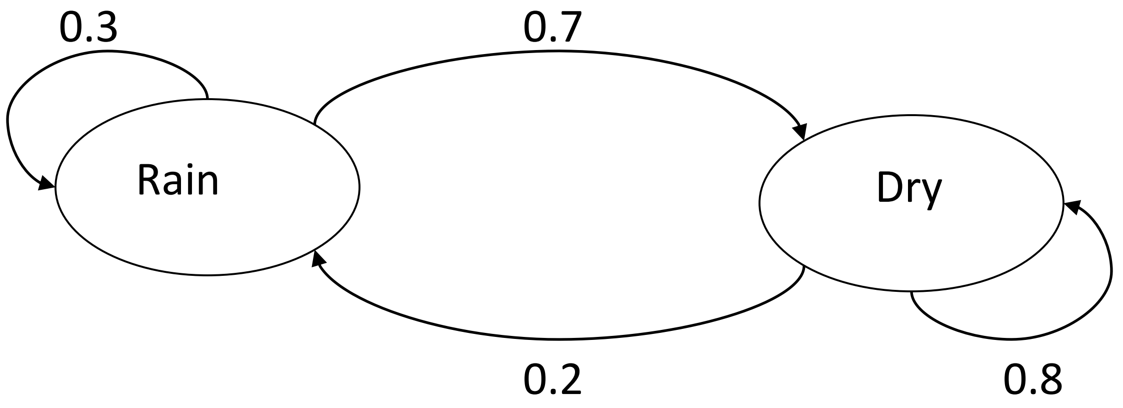
An example of Markov Chain: traffic lights

- A Markov Chain is a sequence of states connected by transitions.

- Traffic lights:
  - 3 States: red, yellow, and green
  - Transition probabilities (0-1):
    - From red to green: P(green|red)=1
    - From green to yellow: P(yellow|green)= 1
    - From yellow to red: P(red|yellow)= 1



Markov Model: two states



- Two states : 'Rain' and 'Dry'.
- Transition probabilities: P('Rain'|'Rain')=0.3 , P('Dry'|'Rain')=0.7 ,
  P('Rain'|'Dry')=0.2, P('Dry'|'Dry')=0.8

- Initial probabilities: say P('Rain')=0.4 , P('Dry')=0.6 .

Calculation of sequence probability

- By Markov chain property, probability of state sequence can be found by the formula:

$$P(s_{i1}, s_{i2}, \ldots, s_{ik}) = P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1}) P(s_{i1}, s_{i2}, \ldots, s_{ik-1})$$

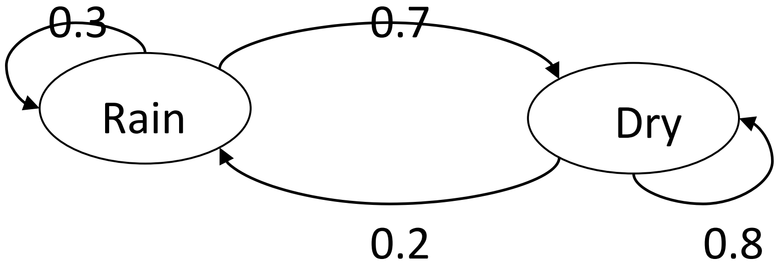
$$= P(s_{ik} \mid s_{ik-1}) P(s_{i1}, s_{i2}, \ldots, s_{ik-1}) = \ldots$$

$$= P(s_{ik} \mid s_{ik-1}) P(s_{ik-1} \mid s_{ik-2}) \ldots P(s_{i2} \mid s_{i1}) P(s_{i1})$$

- Suppose we want to calculate a probability of a sequence of states in our example, {‘Dry’,’Dry’,’Rain’,Rain’}.

P({‘Dry’,’Dry’,’Rain’,Rain’}) =

P(‘Rain’|’Rain’) P(‘Rain’|’Dry’) P(‘Dry’|’Dry’) P(‘Dry’)=



= 0.3*0.2*0.8*0.6 = 0.0288

Calculation of sequence probability

- Given a Markov Chain M where all transition probabilities are known:

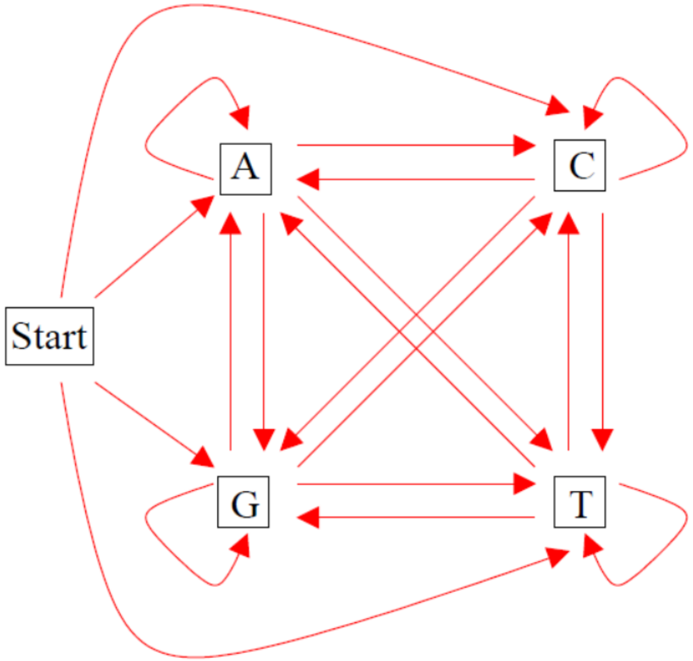
  P(A|G) = 0.18, P(C|G) = 0.38, P(G|G) = 0.32, P(T|G) = 0.12

  P(A|C) = 0.15, P(C|C) = 0.35, P(G|C) = 0.34, P(T|C) = 0.15

  ……

- The probability of sequence x = GCCT is:

  P(GCCT) = P(T|C) * P(C|C) *P(G|C) *P(G)



HMMs are an extension of Markov Chains

- HMMs are like Markov Chains: a finite number of states connected by transitions.
- But the major difference between the two is that the states of a HMM are not a symbol but a set of symbols (observations).
- Each state can emit a symbol (observation) with a probability given by the distribution.

HMMs derive from Markov Models

- Set of states: $\{s_1, s_2, \ldots, s_N\}$
- Process moves from one state to another generating a sequence of states :

$$s_{i1}, s_{i2}, \ldots, s_{ik}, \ldots$$

- Markov chain property: probability of each subsequent state depends only on what was the previous state: $P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1}) = P(s_{ik} \mid s_{ik-1})$

- **States are not visible, but each state randomly generates one of M symbols (observations, visible symbols)** $\{v_1, v_2, \ldots, v_M\}$

- To define hidden Markov model, the following probabilities have to be specified:

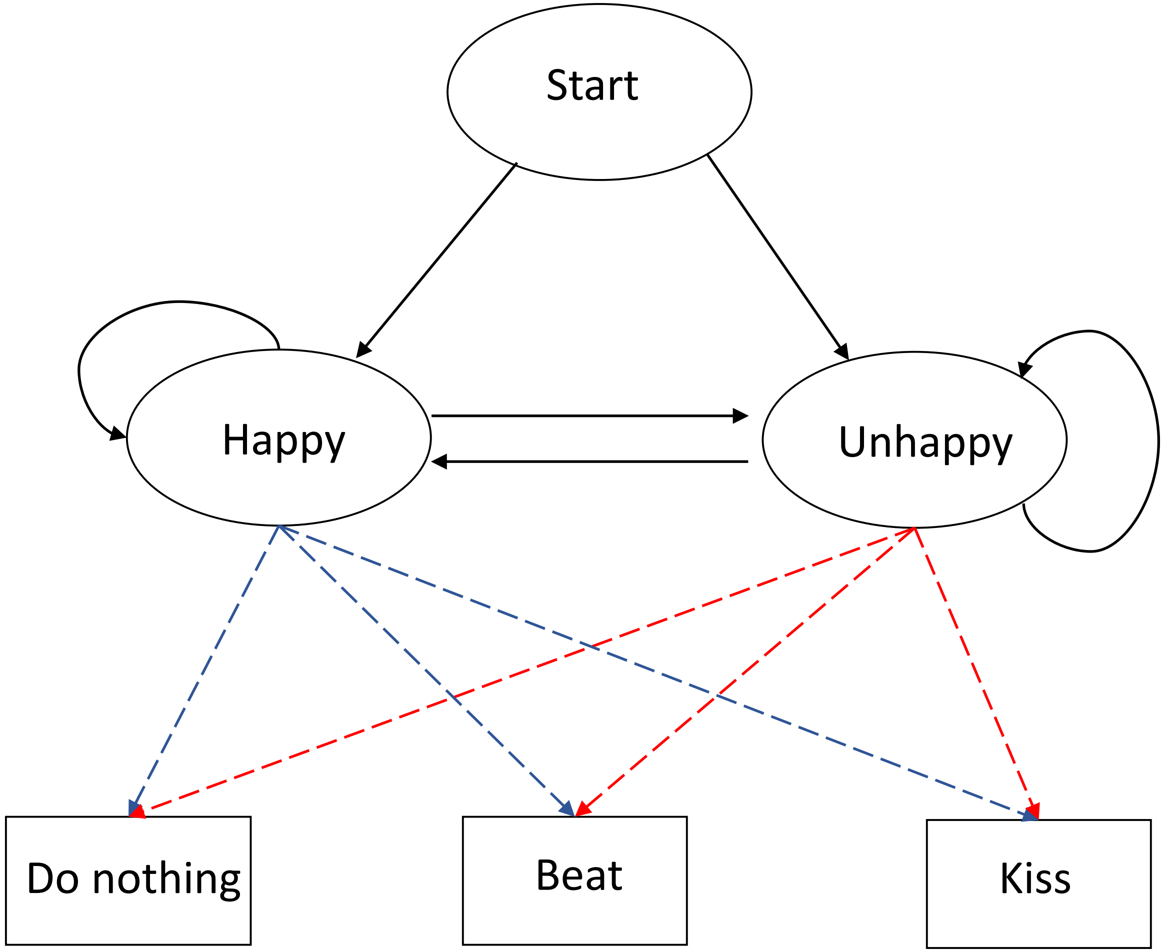
  matrix of transition probabilities A=(a_{ij}), a_{ij}= P(s_i | s_j) ,

  matrix of observation (emission) probabilities B=(b_i (v_m)), b_i(v_m) = P(v_m | s_i)

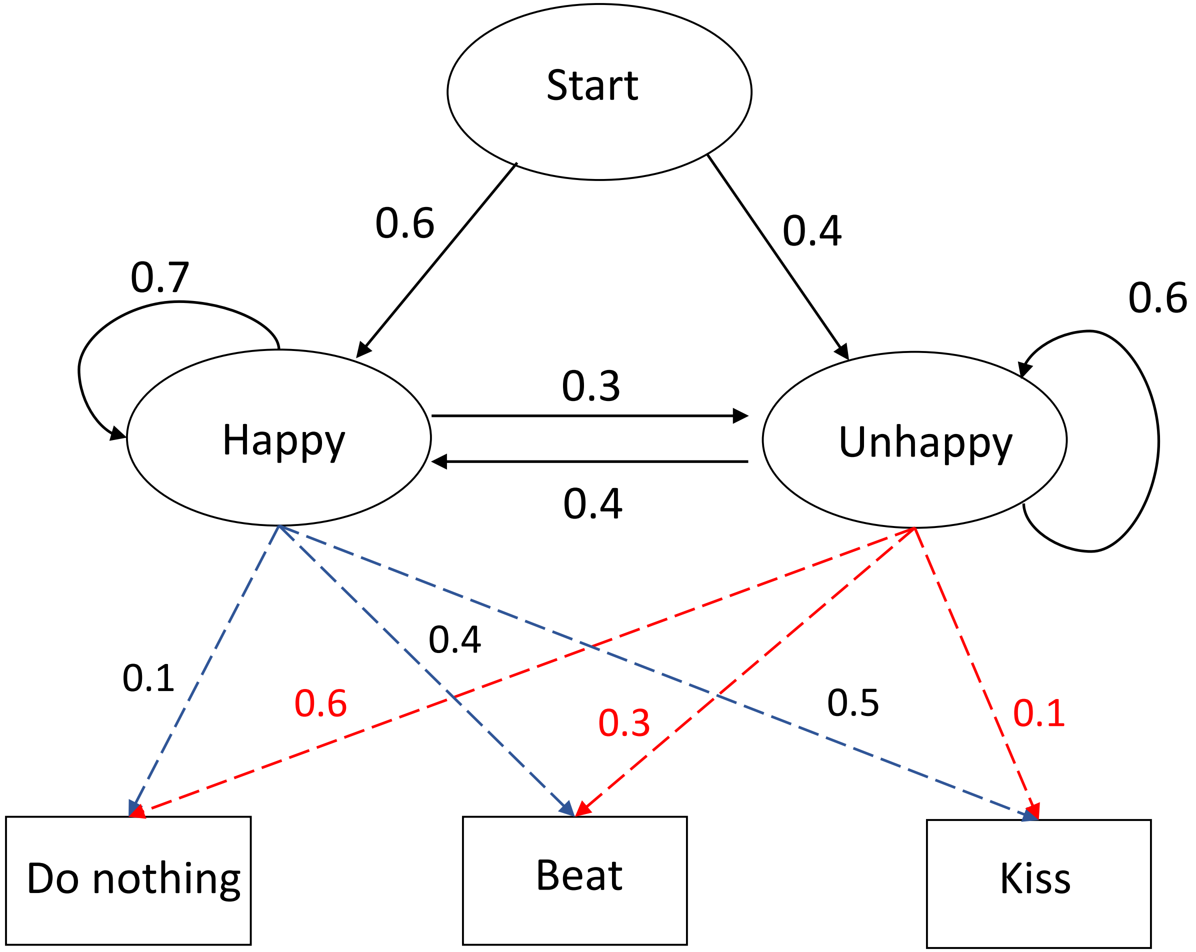
  a vector of initial probabilities π=(π_i), π_i = P(s_i) .

  Model is represented by M=(A, B, π).

Example of HMM: Happy or Unhappy



Example of HMM: Happy or Unhappy

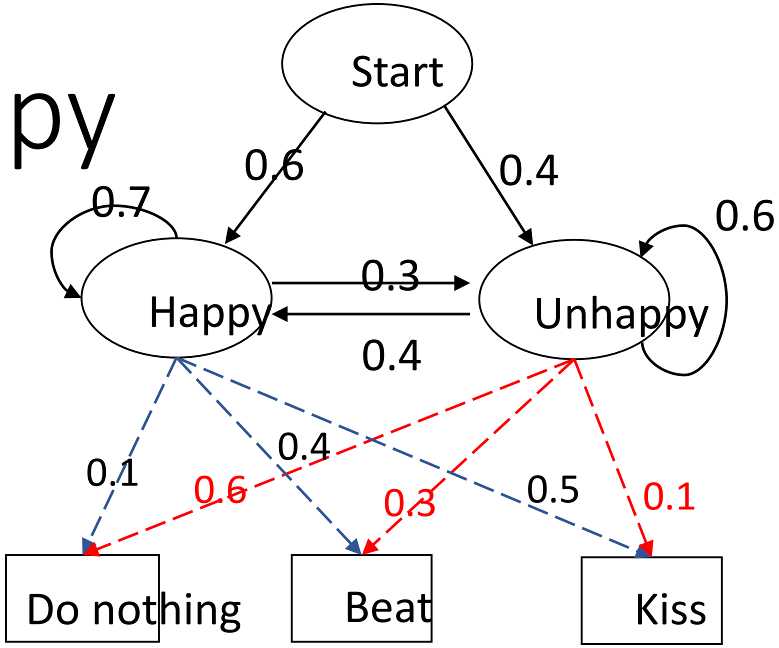


- Hidden states = {Happy, Unhappy}
- Three observations= {Kiss, Beat, Do nothing}

- Initial probabilities={Happy: 0.6; Unhappy: 0.4}

- Transition probabilities= {
  Happy:{Happy: 0.7, Unhappy: 0.3},
  Unhappy: {Happy: 0.4, Unhappy: 0.6},
  }
- Observation (emission) probabilities= {
  Happy:{Kiss: 0.5, Beat: 0.4, Do nothing: 0.1},
  Unhappy: {Kiss: 0.1, Beat: 0.3, Do nothing: 0.6},
  }

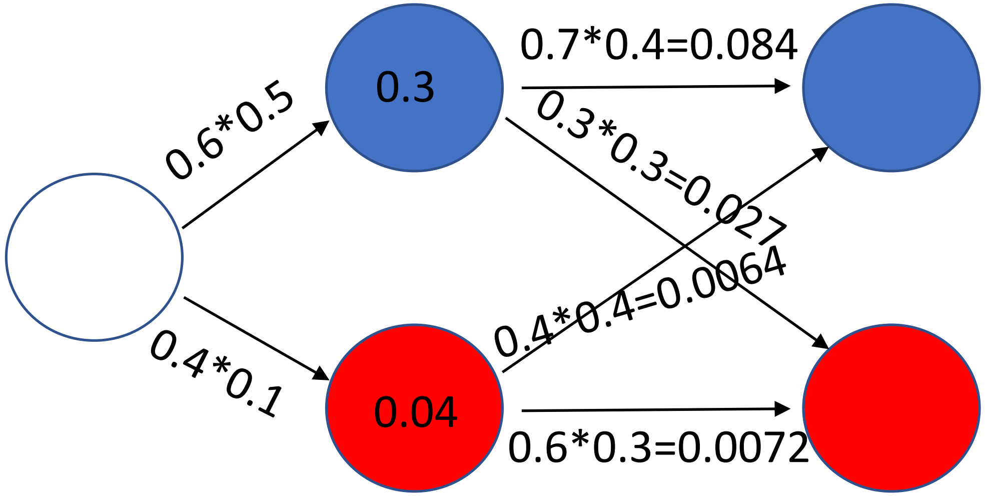
Example of HMM: Happy or Unhappy



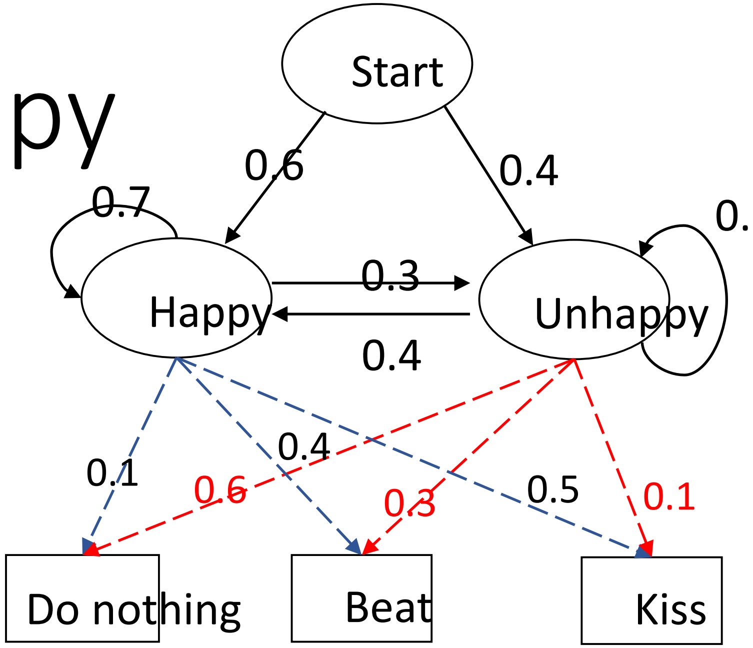
Day 1:
Observation
Kiss

Day 2:
Observation
Beat

Day 3:
Observation
Do nothing



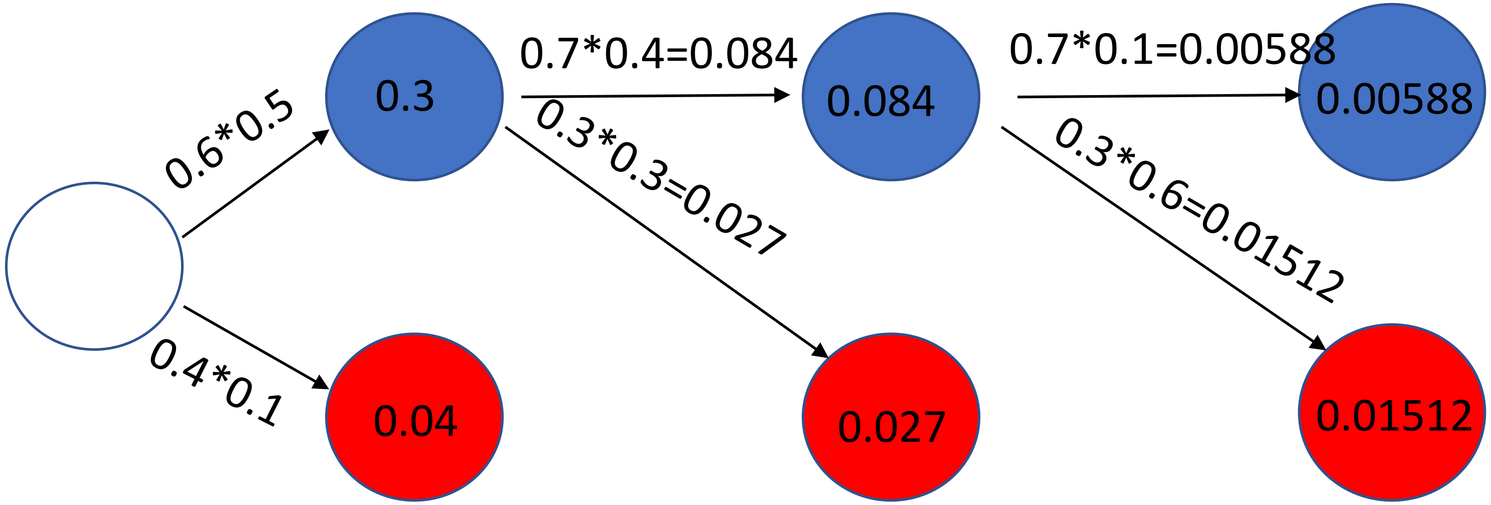
Example of HMM: Happy or Unhappy



Day 1:
Observation
Kiss

Day 2:
Observation
Beat

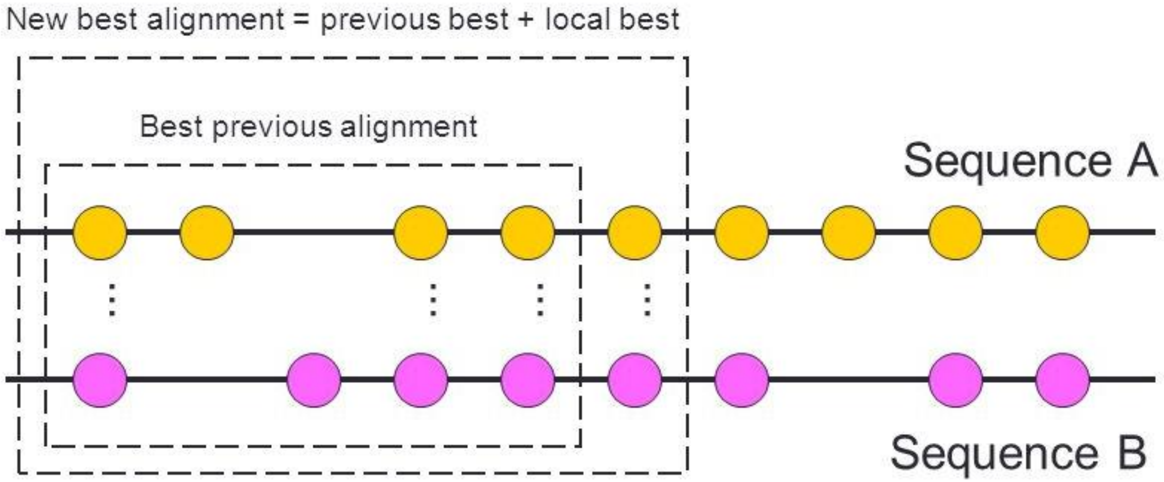
Day 3:
Observation
Do nothing



Outline

- Markov models and HMMs
- Protein profile HMMs
- HMM tools
- Available resources for Profile HMMs

Sequence alignment: a series of states



```
LSP-        LSP-
-TPE        L-PE
  ↓           ↓
XMMY        MXMY
```

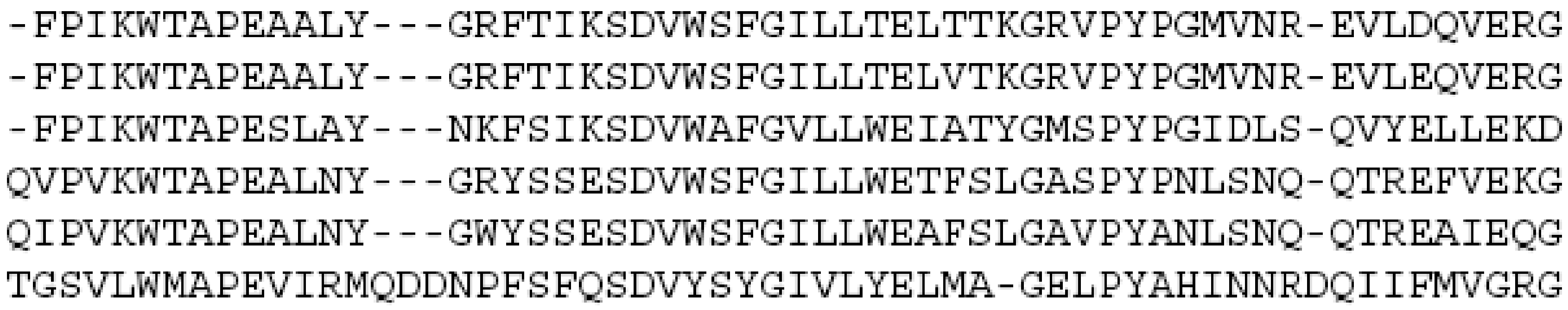
M: Match (not necessarily identical)

X: Insert at sequence X
(delete at sequence Y)

Y: Insert at sequence Y
(delete at sequence X)

Protein family and Profile Hidden Markov Models

- Multiple sequence alignment



```
-FPIKWTAPEAALY---GRFTIKSDVWSFGILLTELTTKGRVPYPGMVNR-EVLDQVERG
-FPIKWTAPEAALY---GRFTIKSDVWSFGILLTELVTKGRVPYPGMVNR-EVLEQVERG
-FPIKWTAPESLAY---NKFSIKSDVWAFGVLLWEIATYGMSPYPGIDLS-QVYELLEKD
QVPVKWTAPEALNY---GRYSSESDVWSFGILLWETFSLGASPYPNLSNQ-QTREFVEKG
QIPVKWTAPEALNY---GWYSSESDVWSFGILLWEAFSLGAVPYANLSNQ-QTREAIEQG
TGSVLWMAPEVIRMQDDNPFSFQSDVYSYGIVLYELMA-GELPYAHINNRDQIIFMVGRG
```

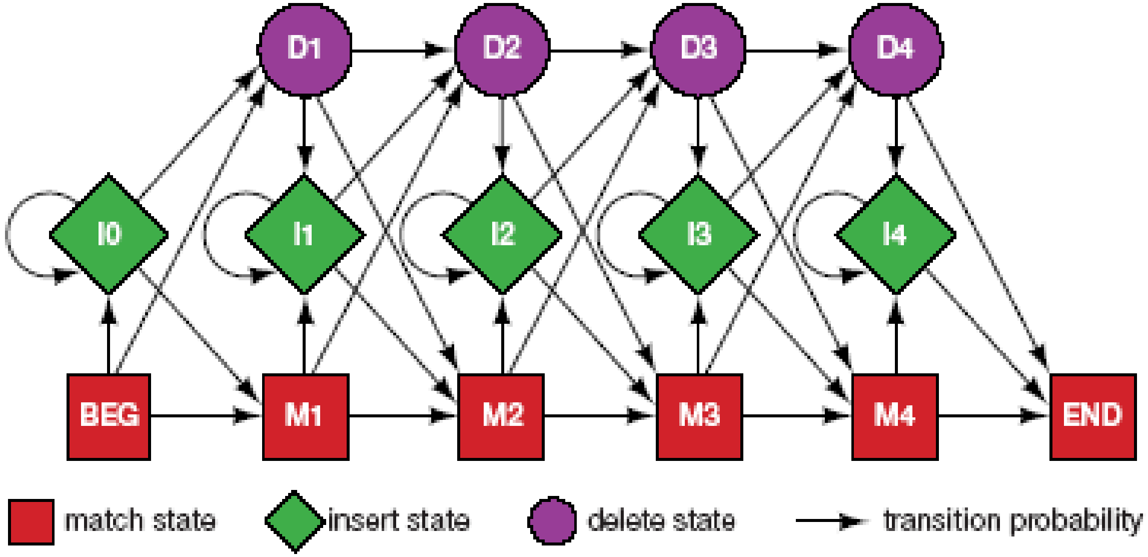
  - Each consensus column can exist in 3 states:
  - Match, Insert and Delete states

  - Number of states depends upon length of the alignment

  - Each Match state generates one of 20 amino acids (symbols or observations)

Profile Hidden Markov Models

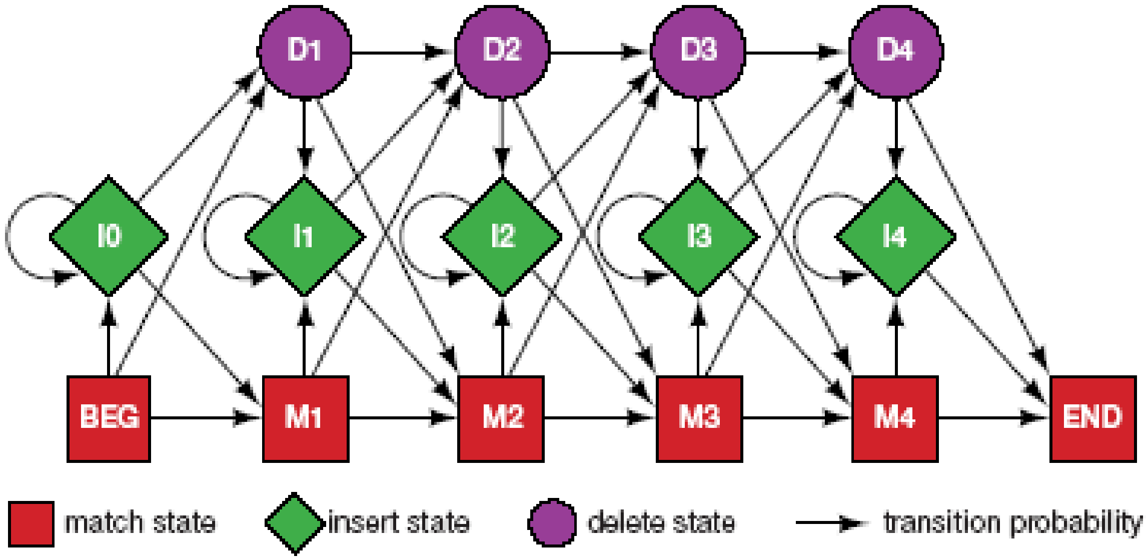
- A typical profile HMM architecture



- Squares represent match states
- Diamonds represent insert states
- Circles represent delete states
- Arrows represent transitions

Profile Hidden Markov Models

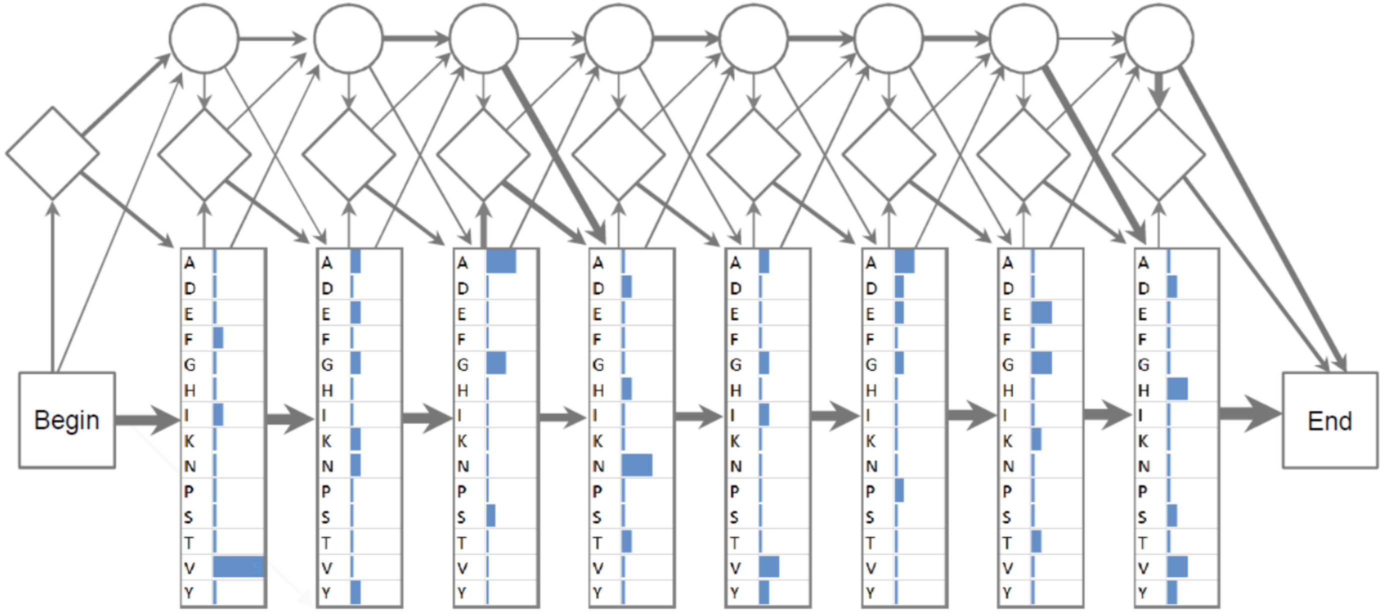
- Estimation of parameters



- transition probabilities estimated as frequency of a transition in a given alignment
- emission probabilities estimated as frequency of an emission (symbol) in a given alignment
- pseudo counts usually introduced to account for transititions / emissions which were not present in the alignment

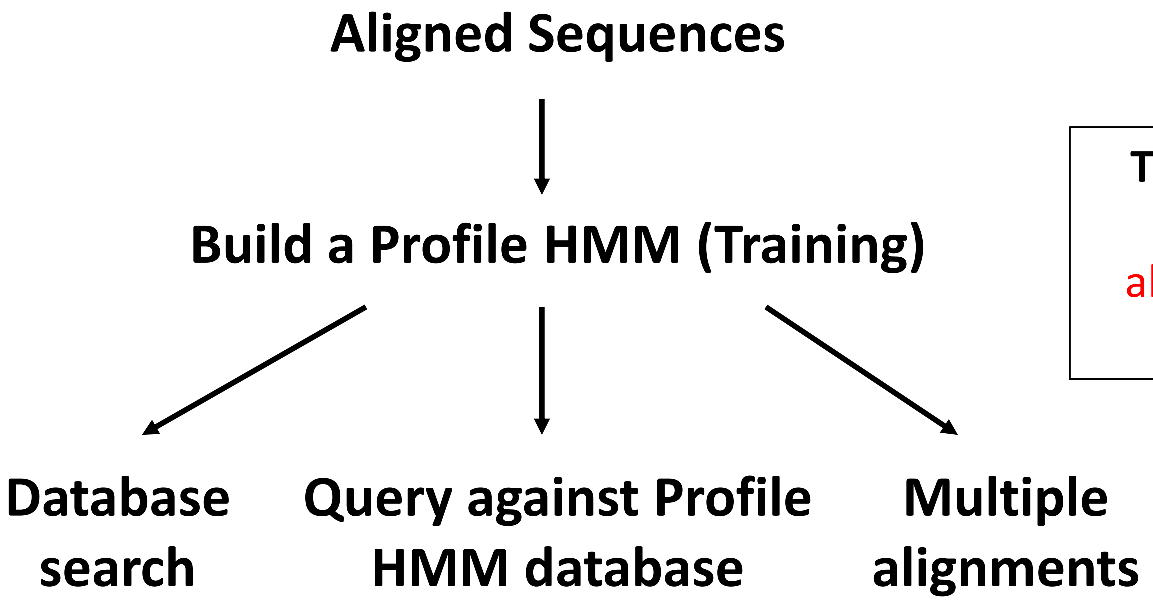
In the case of a “real” alignment, an HMM for it might look like this:

```
HBA_HUMAN      ...VGA--HAGEY...
HBB_HUMAN      ...V----NVDEV...
MYG_PHYCA      ...VEA--DVAGH...
GLB3_CHITP     ...VKG------D...
GLB5_PETMA     ...VYS--TYETS...
LGB2_LUPLU     ...FNA--NIPKH...
GLB1_GLYDI     ...IAGADNGAGV...
                  ***  *****
```



Begin
End
A D E F G H I K N P S T V Y
A D E F G H I K N P S T V Y
A D E F G H I K N P S T V Y
A D E F G H I K N P S T V Y
A D E F G H I K N P S T V Y
A D E F G H I K N P S T V Y
A D E F G H I K N P S T V Y
A D E F G H I K N P S T V Y

Three important questions can be answered using HMMs



Aligned Sequences

Build a Profile HMM (Training)

Database search

Query against Profile HMM database

Multiple alignments

Training problem and is solved using the Forward-backward algorithm and the Baum-Welch expectation maximization

Scoring problem and it can be solved using the Forward algorithm

Alignment problem and it is solved by the Viterbi algorithm

For details about these algorithms see: Durbin, Eddy, Mitchison, Krog. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1998.

Outline

- Markov models and HMMs
- Protein profile HMMs
- HMM tools
- Available resources for Profile HMMs

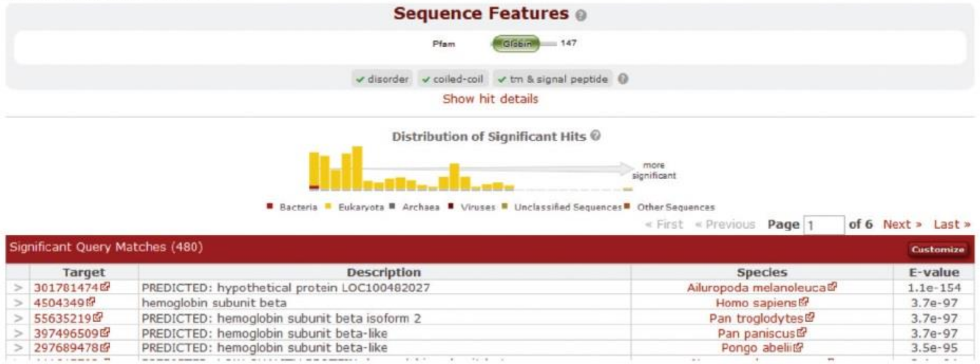
HMMs: Tools

- **HMMER3** is a package to build and use HMMs developed by Sean Eddy (http://hmmer.wustl.edu/).
- Software available in HMMER2:
  - hmmbuild to build an HMM from a multiple alignment;
  - hmmalign to align sequences to an HMM model;
  - hmmsearch to search a sequence database with an HMM model;
  - jackhmmer to iteratively search sequence(s) against a protein database;
  - hmmscan to search protein sequence(s) against a protein profile database;
  - hmmemit to get sample sequences from a profile HMM;
  - hmmfetch to retrieve profile HMM(s) from a file

- **SAM** is a similar package developed by Richard Hughey, Kevin Karplus and Anders Krogh (http://www.cse.ucsc.edu/research/compbio/sam.html).

HMMER is available online

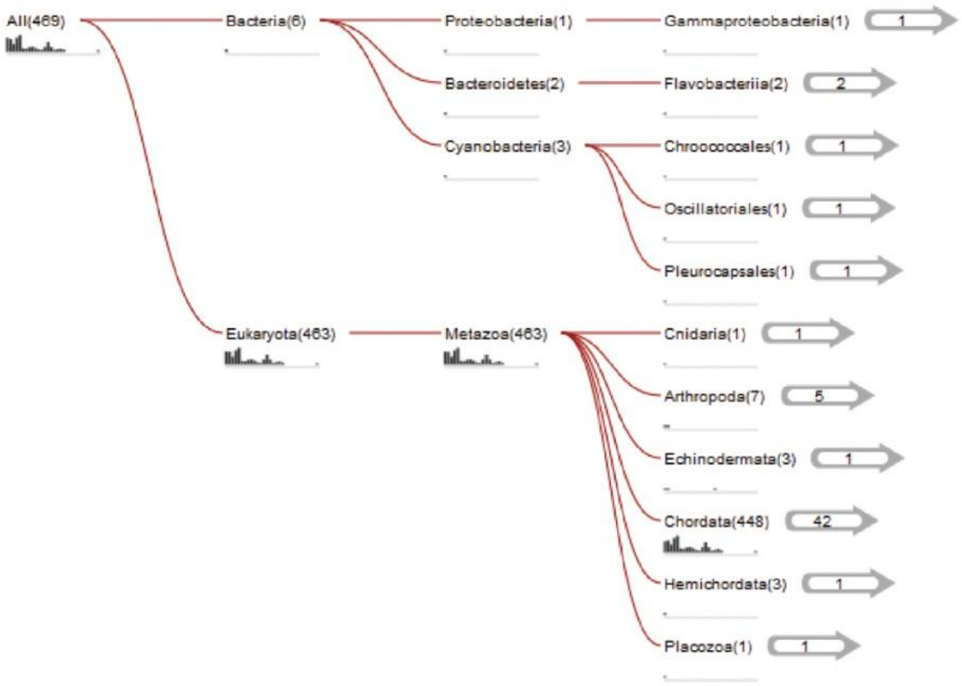
(a) HMMER web output



Significant Query Matches (480)

	Target	Description	Species	E-value
>	301781474	PREDICTED: hypothetical protein LOC100482027	Ailuropoda melanoleuca	1.1e-154
>	4504349	hemoglobin subunit beta	Homo sapiens	3.7e-97
>	55635219	PREDICTED: hemoglobin subunit beta isoform 2	Pan troglodytes	3.7e-97
>	397496509	PREDICTED: hemoglobin subunit beta-like	Pan paniscus	3.7e-97
>	297689478	PREDICTED: hemoglobin subunit beta-like	Pongo abelii	3.5e-95

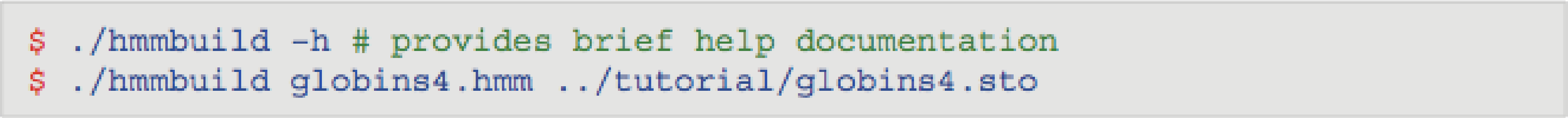
(b) HMMER phylogenetic output



- Domain architecture
- Taxonomy
- Iterative manner

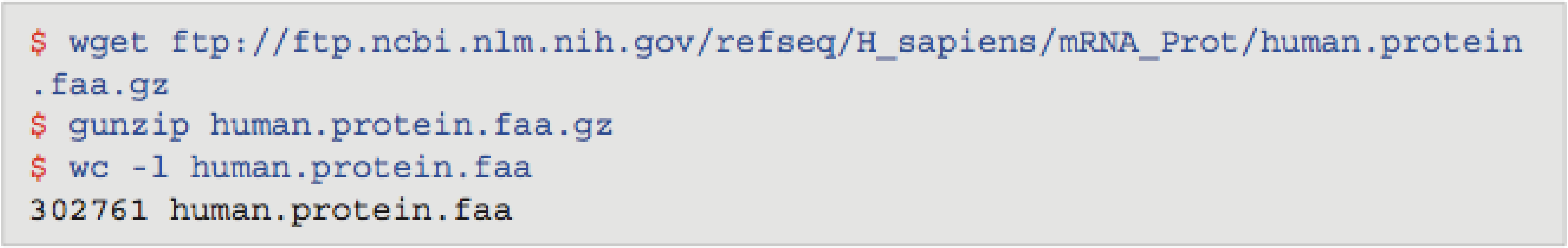
HMMER software: build profiles, complement BLAST

Build a profile HMM (input is a multiple sequence alignment)



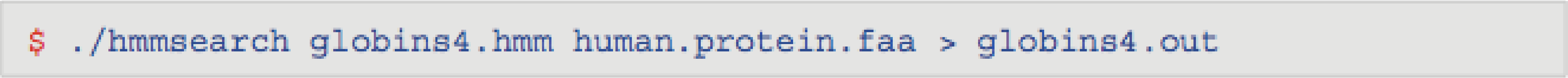
```
$ ./hmmbuild -h # provides brief help documentation
$ ./hmmbuild globins4.hmm ../tutorial/globins4.sto
```

Download a database to search (e.g. human RefSeq proteins)



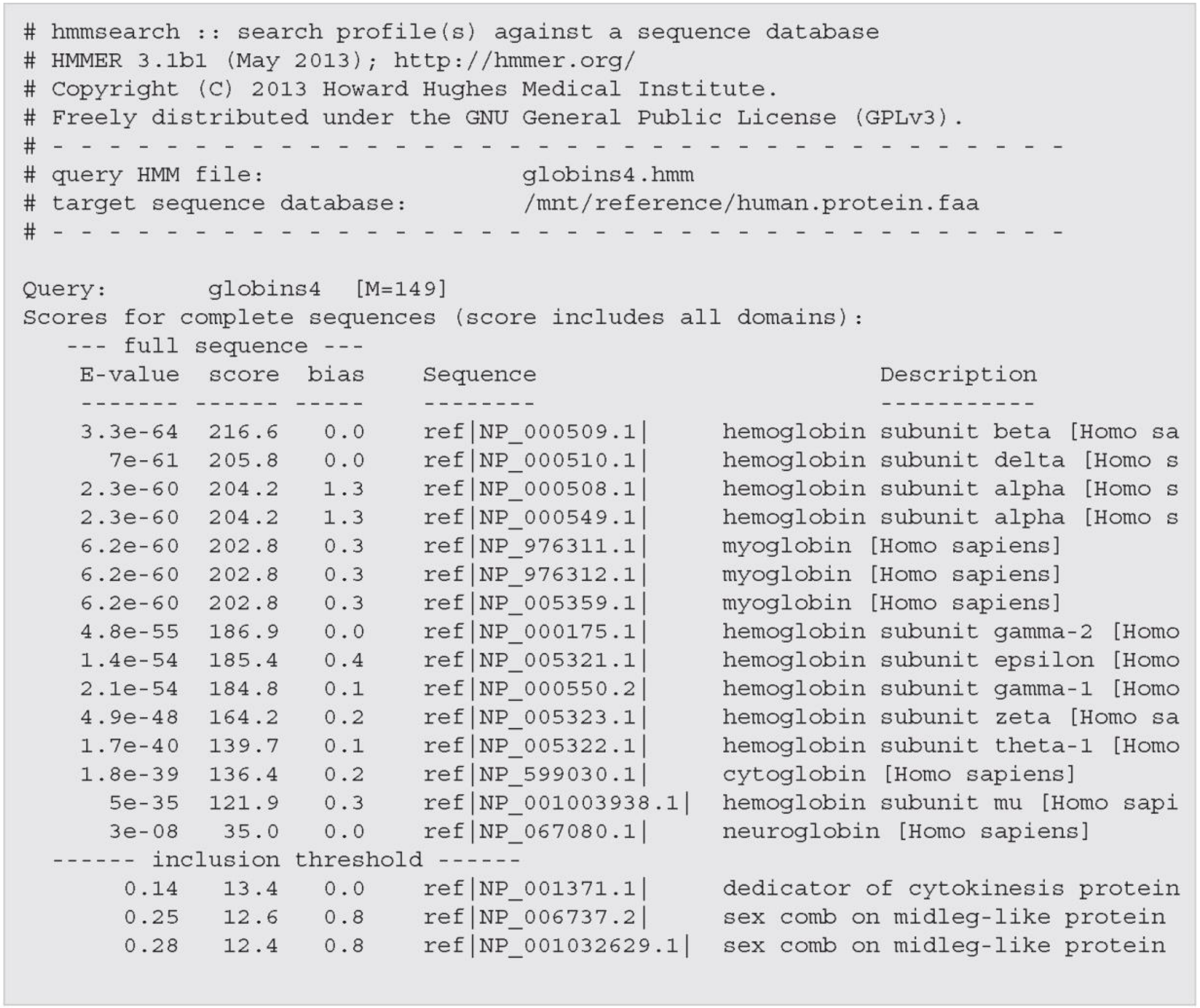
```
$ wget ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein
.faa.gz
$ gunzip human.protein.faa.gz
$ wc -l human.protein.faa
302761 human.protein.faa
```

Search an HMM against a database



```
$ ./hmmsearch globins4.hmm human.protein.faa > globins4.out
```

HMMER results



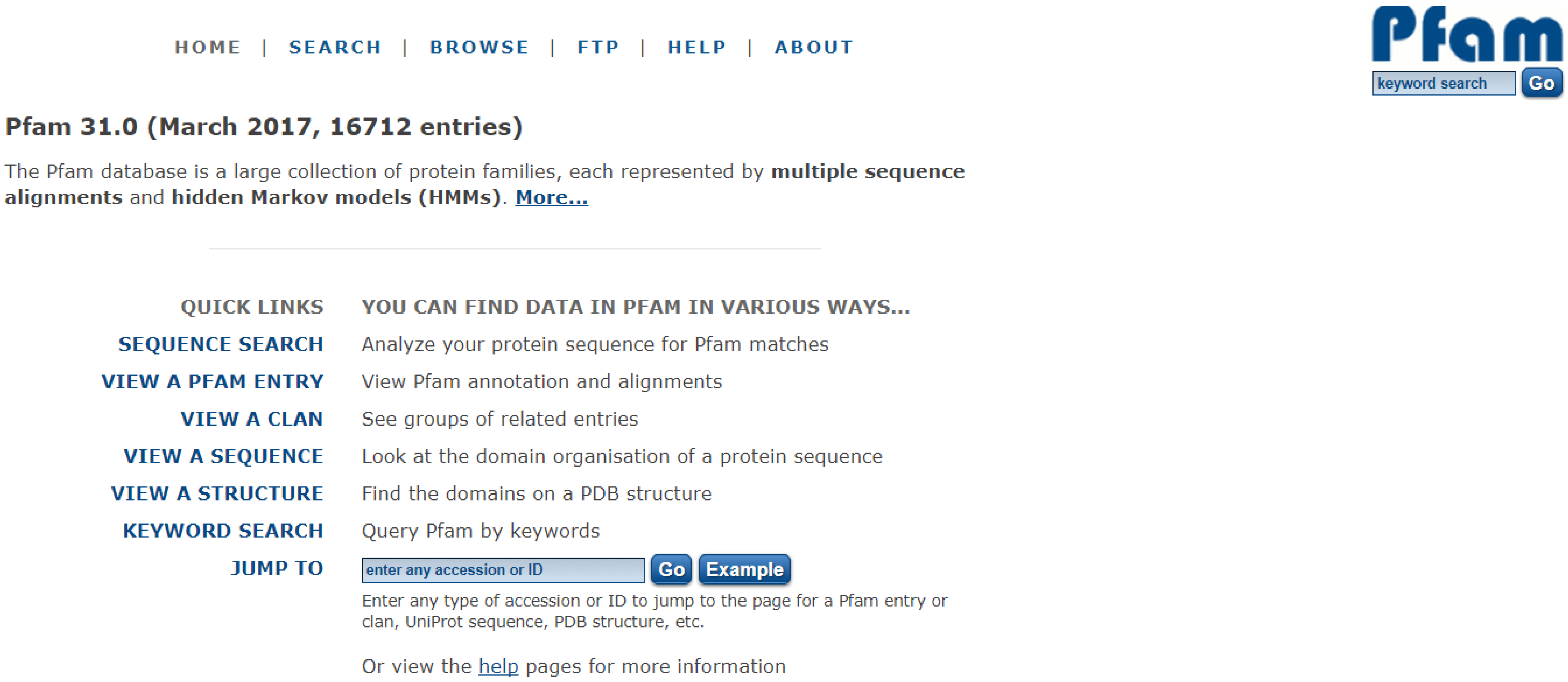
```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b1 (May 2013); http://hmmer.org/
# Copyright (C) 2013 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  globins4.hmm
# target sequence database:        /mnt/reference/human.protein.faa
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       globins4  [M=149]
Scores for complete sequences (score includes all domains):
   --- full sequence ---
    E-value  score  bias    Sequence                        Description
    ------- ------ -----    --------                        -----------
    3.3e-64  216.6    0.0    ref|NP_000509.1|     hemoglobin subunit beta [Homo sa
      7e-61  205.8    0.0    ref|NP_000510.1|     hemoglobin subunit delta [Homo s
    2.3e-60  204.2    1.3    ref|NP_000508.1|     hemoglobin subunit alpha [Homo s
    2.3e-60  204.2    1.3    ref|NP_000549.1|     hemoglobin subunit alpha [Homo s
    6.2e-60  202.8    0.3    ref|NP_976311.1|     myoglobin [Homo sapiens]
    6.2e-60  202.8    0.3    ref|NP_976312.1|     myoglobin [Homo sapiens]
    6.2e-60  202.8    0.3    ref|NP_005359.1|     myoglobin [Homo sapiens]
    4.8e-55  186.9    0.0    ref|NP_000175.1|     hemoglobin subunit gamma-2 [Homo
    1.4e-54  185.4    0.4    ref|NP_005321.1|     hemoglobin subunit epsilon [Homo
    2.1e-54  184.8    0.1    ref|NP_000550.2|     hemoglobin subunit gamma-1 [Homo
    4.9e-48  164.2    0.2    ref|NP_005323.1|     hemoglobin subunit zeta [Homo sa
    1.7e-40  139.7    0.1    ref|NP_005322.1|     hemoglobin subunit theta-1 [Homo
    1.8e-39  136.4    0.2    ref|NP_599030.1|     cytoglobin [Homo sapiens]
      5e-35  121.9    0.3    ref|NP_001003938.1|  hemoglobin subunit mu [Homo sapi
      3e-08   35.0    0.0    ref|NP_067080.1|     neuroglobin [Homo sapiens]
  ------ inclusion threshold ------
       0.14   13.4    0.0    ref|NP_001371.1|     dedicator of cytokinesis protein
       0.25   12.6    0.8    ref|NP_006737.2|     sex comb on midleg-like protein
       0.28   12.4    0.8    ref|NP_001032629.1|  sex comb on midleg-like protein
```

HMMER output includes scores, E values

Outline

- Markov models and HMMs
- Protein profile HMMs
- HMM tools
- Available resources for Profile HMMs

Pfam--- Protein Domain database (pfam.xfam.org/)



Pfam 31.0 (March 2017, 16712 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

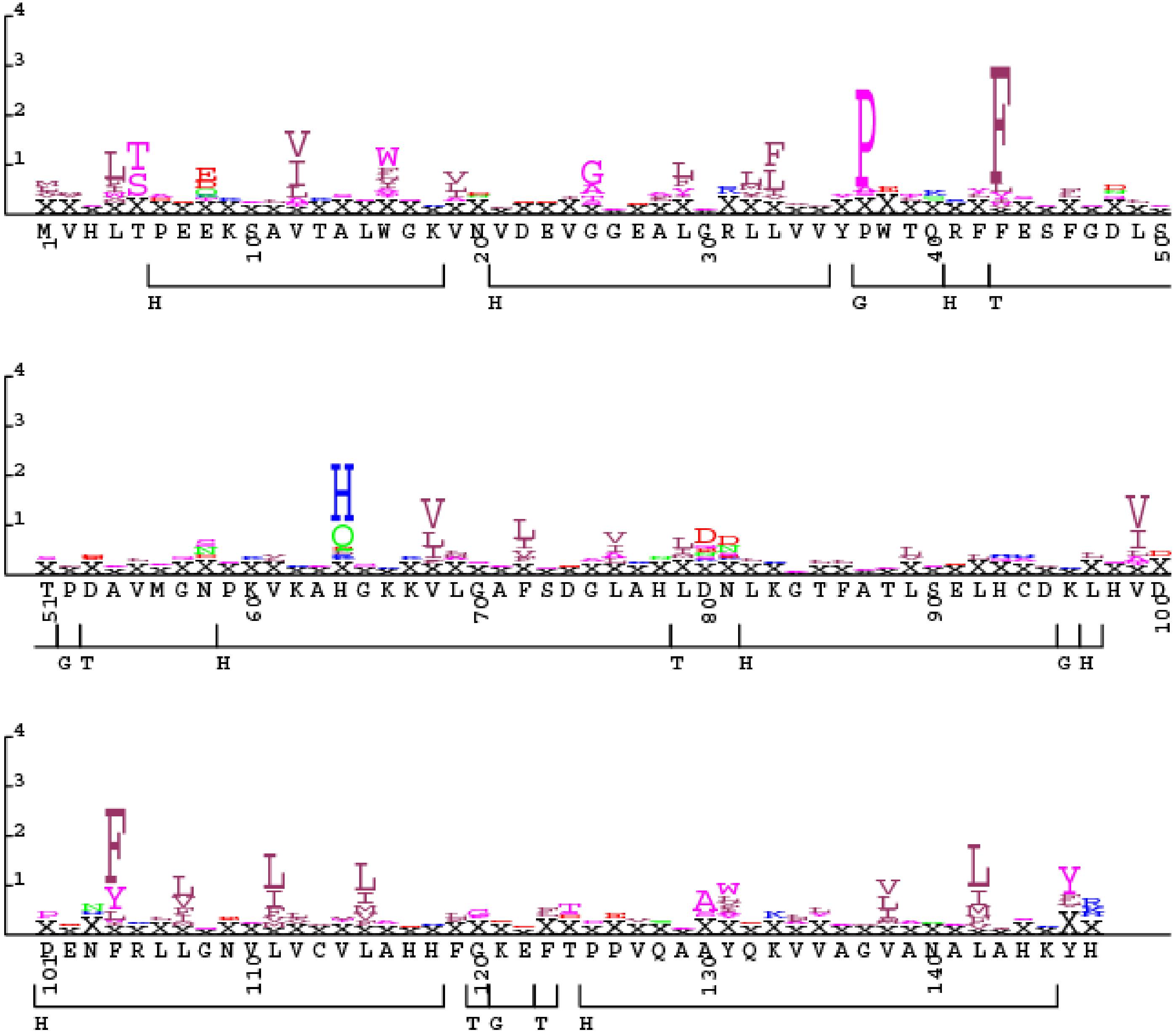
Pfam--- Protein Domain database (pfam.xfam.org/)

- Pfam is a database of multiple alignments and hidden Markov models (HMMs) of common conserved protein domains.

- The alignments use a non-redundant protein set composed of SWISS-PROT and TrEMBL.

- Pfam consists of parts A and B.
    - Pfam-A contains curated domain families with high-quality alignments.
    - Pfam-B contains families that were generated automatically by clustering the remaining sequences after removal of Pfam-A domains.

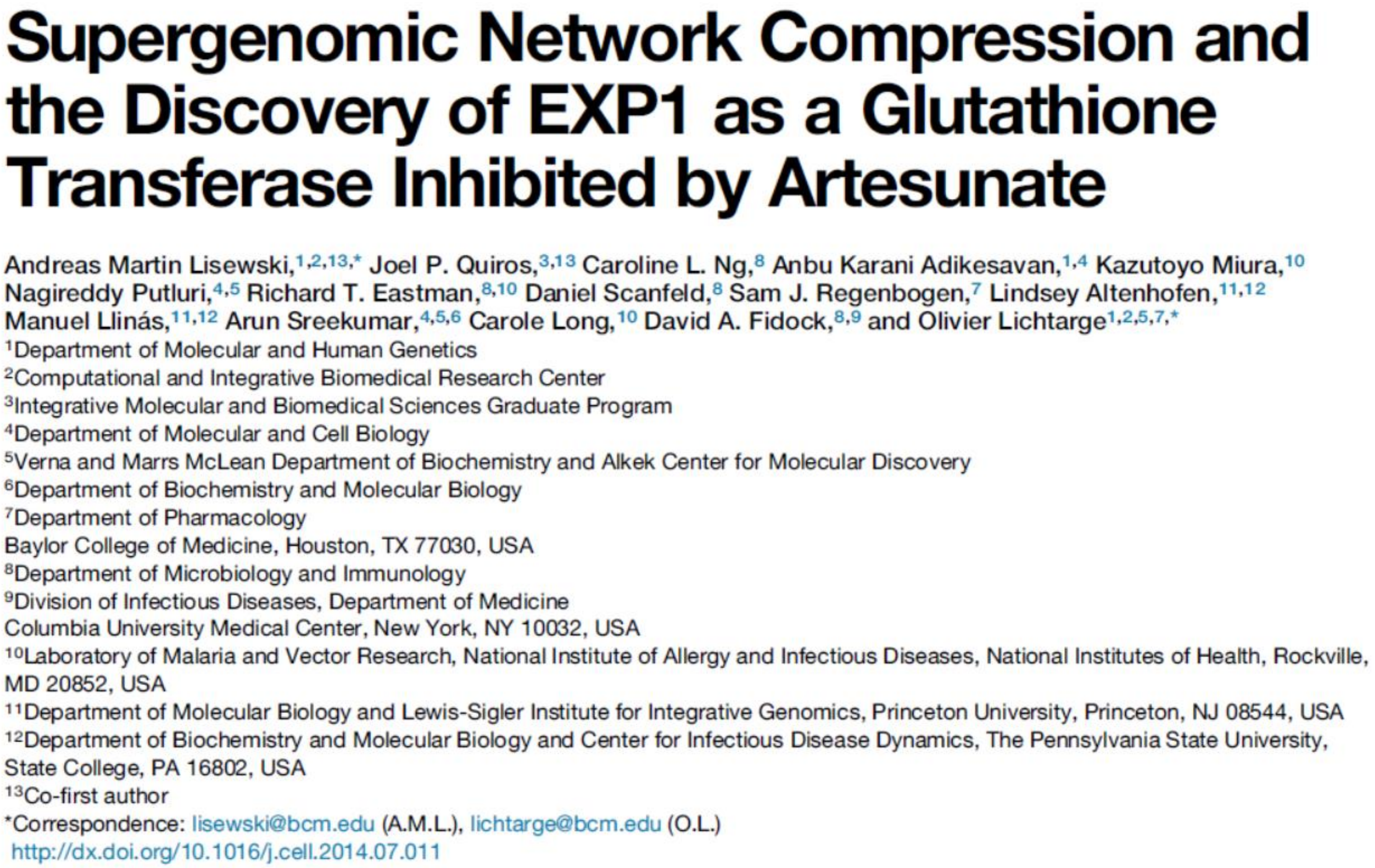
HMM logos graphically depict the likelihood of observed amino acids at Pfam



HMM *versus* PSSM

- **Advantages:**
  - A HMM has position-dependent amino acid distributions, which are represented as emission probabilities at each match state. (also PSSM)
  - Insertion/deletion gap penalties are handled using transition probabilities. (Usually not with PSSM)
  - The possible dependence of an amino acid on its preceding neighbor can be represented using the transition probabilities. (Not with PSSM)

- **Problems:**
  - Long-range interactions between amino acids.
  - Requirement of multiple sequence alignments.

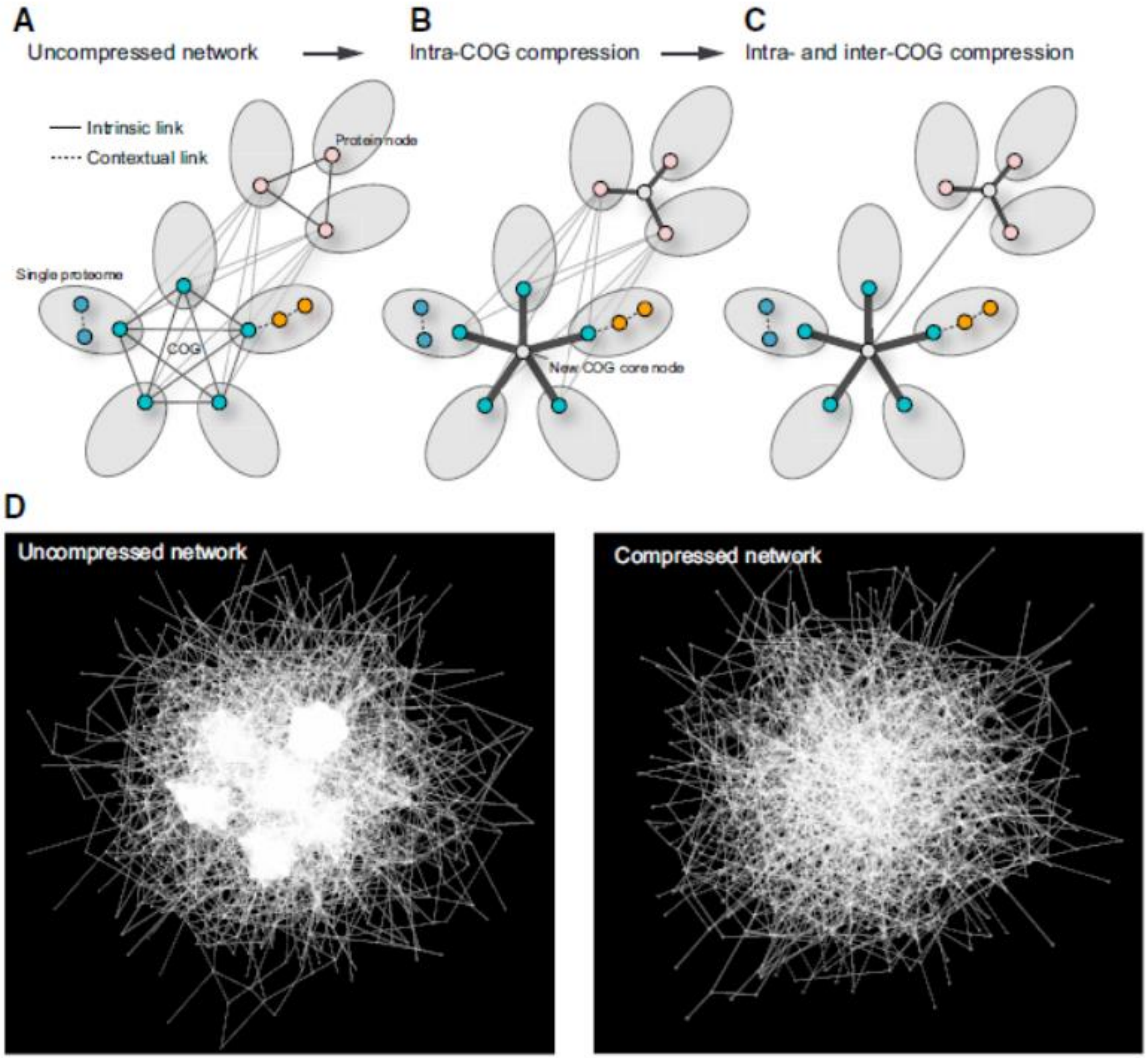


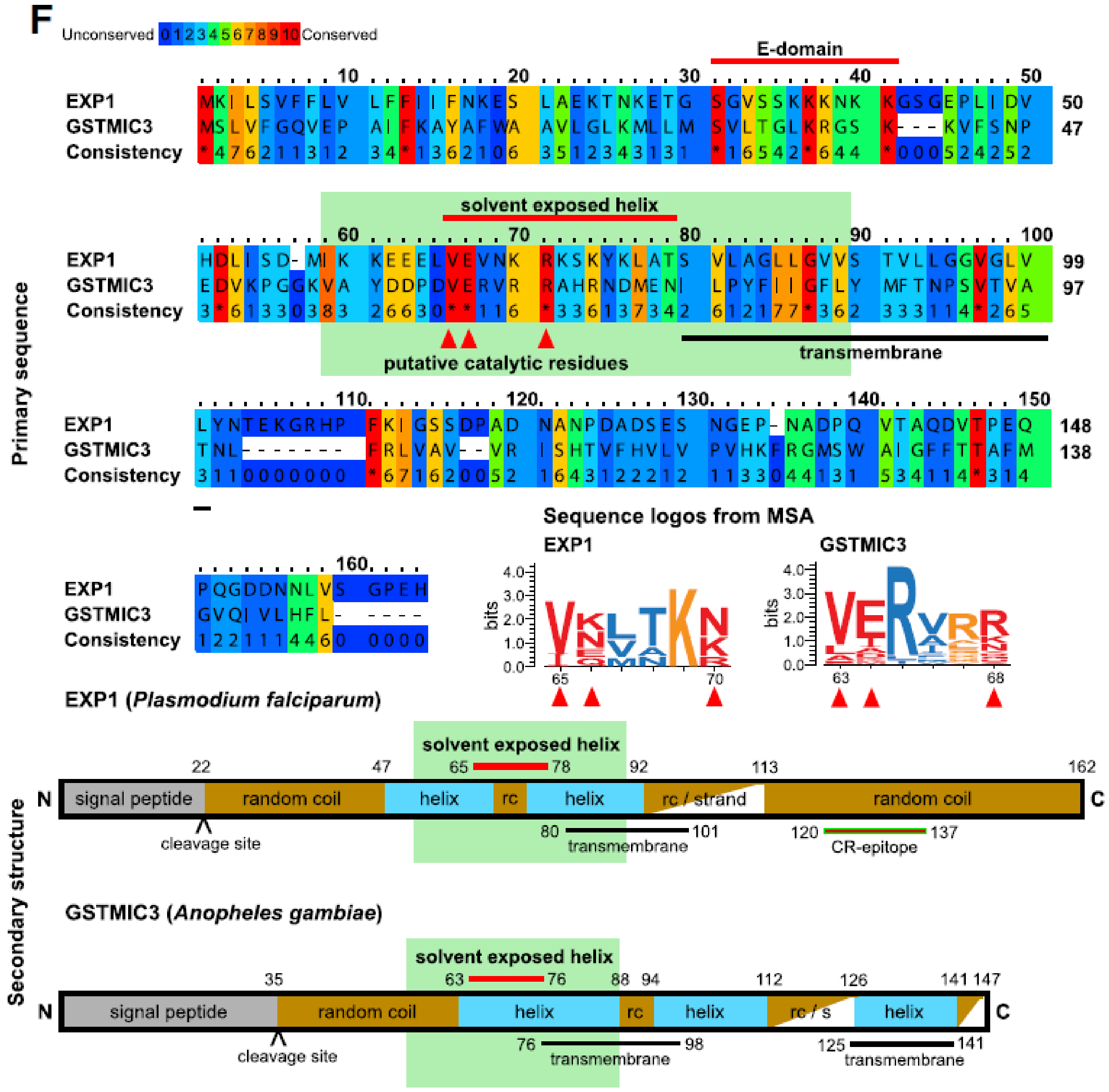
Supergenomic Network Compression and the Discovery of EXP1 as a Glutathione Transferase Inhibited by Artesunate

Andreas Martin Lisewski,[1,2,13,*] Joel P. Quiros,[3,13] Caroline L. Ng,[8] Anbu Karani Adikesavan,[1,4] Kazutoyo Miura,[10] Nagireddy Putluri,[4,5] Richard T. Eastman,[8,10] Daniel Scanfeld,[8] Sam J. Regenbogen,[7] Lindsey Altenhofen,[11,12] Manuel Llinás,[11,12] Arun Sreekumar,[4,5,6] Carole Long,[10] David A. Fidock,[8,9] and Olivier Lichtarge[1,2,5,7,*]

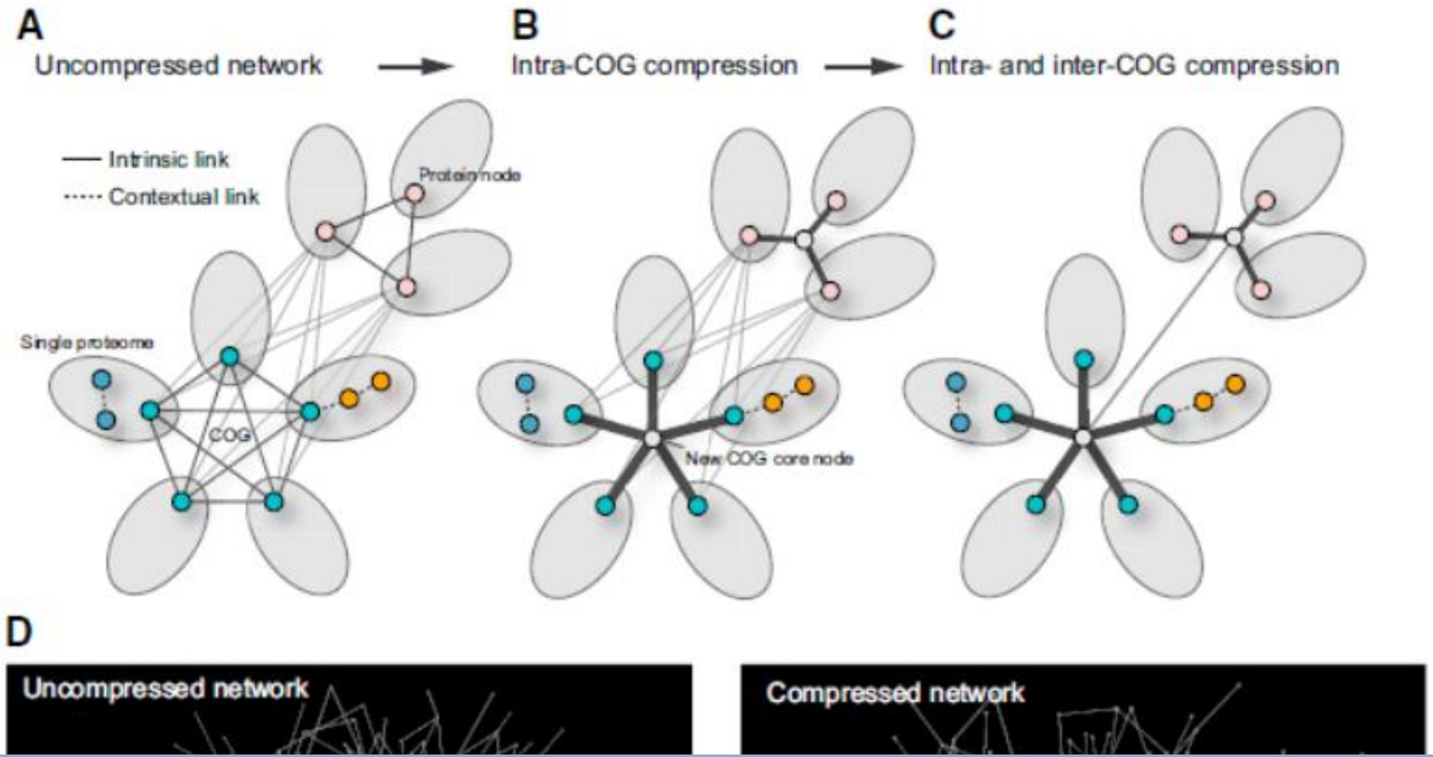
[1]Department of Molecular and Human Genetics
[2]Computational and Integrative Biomedical Research Center
[3]Integrative Molecular and Biomedical Sciences Graduate Program
[4]Department of Molecular and Cell Biology
[5]Verna and Marrs McLean Department of Biochemistry and Alkek Center for Molecular Discovery
[6]Department of Biochemistry and Molecular Biology
[7]Department of Pharmacology
Baylor College of Medicine, Houston, TX 77030, USA
[8]Department of Microbiology and Immunology
[9]Division of Infectious Diseases, Department of Medicine
Columbia University Medical Center, New York, NY 10032, USA
[10]Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD 20852, USA
[11]Department of Molecular Biology and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
[12]Department of Biochemistry and Molecular Biology and Center for Infectious Disease Dynamics, The Pennsylvania State University, State College, PA 16802, USA
[13]Co-first author
*Correspondence: lisewski@bcm.edu (A.M.L.), lichtarge@bcm.edu (O.L.)

http://dx.doi.org/10.1016/j.cell.2014.07.011


A
Uncompressed network
B
Intra-COG compression
C
Intra- and inter-COG compression
Intrinsic link
Contextual link
Protein node
Single proteome
COG
New COG core node
D
Uncompressed network
Compressed network


F
Unconserved 0 1 2 3 4 5 6 7 8 9 10 Conserved
E-domain
Primary sequence
10 20 30 40 50
EXP1 MKILSVFFLV LFFIIFNKES LAEKTNKETG SGVSSKKKNK KGSGEPLIDV 50
GSTMIC3 MSLVFGQVEP AIFKAYAFWA AVLGLKMLLM SVLTGLKRGS K---KVFSNP 47
Consistency *476211312 34*1362106 3512343131 *16542*644 *000524252
solvent exposed helix
60 70 80 90 100
EXP1 HDLISD-MIK KEEELVEVNK RKSKYKLATS VLAGLLGVVS TVLLGGVGLV 99
GSTMIC3 EDVKPGGKVA YDDPDVERVR RAHRNDMENI LPYFIIGFLY MFTNPSVTVA 97
Consistency 3*61330383 26630**116 *336137342 6121777*362 333114*265
putative catalytic residues
transmembrane
110 120 130 140 150
EXP1 LYNTEKGRHP FKIGSSDPAD NANPDADSES NGEP-NADPQ VTAQDVTPEQ 148
GSTMIC3 TNL------- FRLVAV--VR ISHTVFHVLV PVHKFRGMSW AIGFFTTAFM 138
Consistency 3110000000 *671620052 1643122212 1133044131 534114*314
160
EXP1 PQGDDNNLVS GPEH
GSTMIC3 GVQIVLHFL- ----
Consistency 1221114460 0000
Sequence logos from MSA
EXP1
GSTMIC3
bits
4.0 3.0 2.0 1.0 0.0
65 70
63 68
EXP1 (Plasmodium falciparum)
Secondary structure
solvent exposed helix
22 47 65 78 92 113 162
N signal peptide random coil helix rc helix rc / strand random coil C
cleavage site
80 101 transmembrane
120 137 CR-epitope
GSTMIC3 (Anopheles gambiae)
solvent exposed helix
35 63 76 88 94 112 126 141 147
N signal peptide random coil helix rc helix rc / s helix C
cleavage site
76 98 transmembrane
125 141 transmembrane

1. BLAST/PSI-BLAST→ multiple sequence alignment
2. RPS-BLAST→ CDD
3. HMMER→ Pfam
4. Structural information



1. Sequence/structure information is the most reliable (and only) source for homology
2. Current COG information, although provides homology information, didn’t show any better performance than PSI-BLAST
3. Contextual information (PPI, genomic association, co-expression) indicates the functional relevance not the homology
4. Gene Ontology: false annotation, low coverage