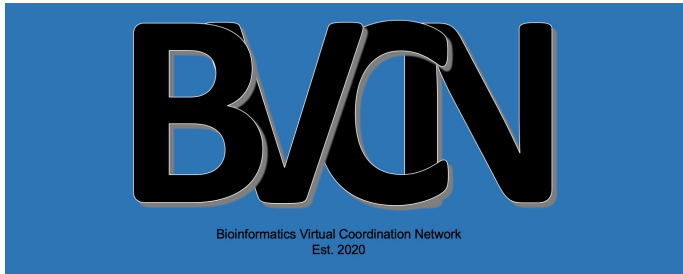
Functional Annotation Lesson 3

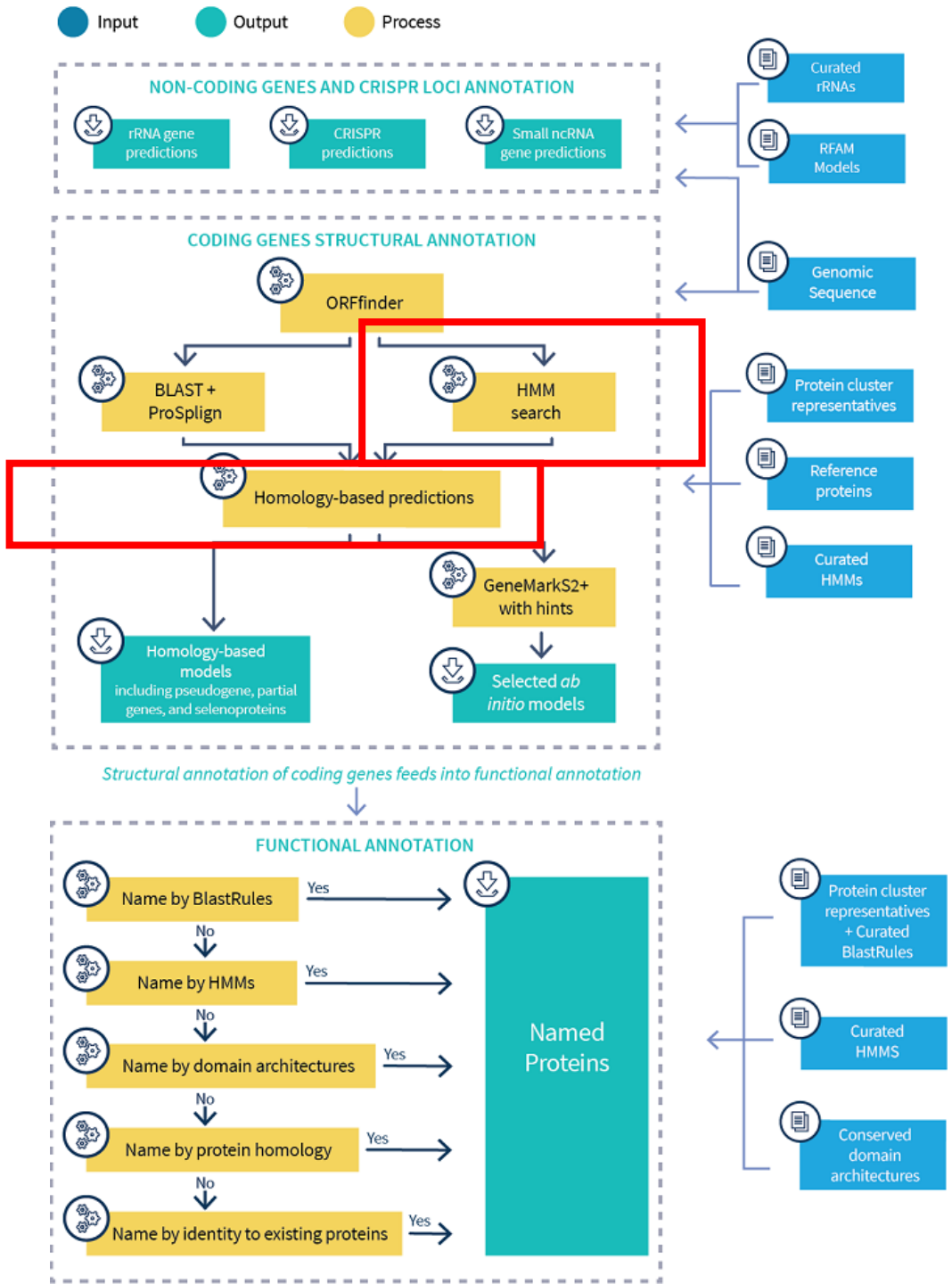
Inferring function from position-sensitive models

- HMMER



NCBI PGAP

Prokaryotic Genome Annotation Pipeline



The basis for the prediction of features is nearly always a sequence alignment

Based on experimentally verified sequence annotations, a multiple sequence alignment is constructed

Different methods exist to capture the information gained from this multiple sequence alignment

PROBABILISTIC APPROACHES to Homology

- BLAST is computationally greedy, alternative annotation options can use a probabilistic approach
- Probabilistic approach – incorporate random variables and probability distributions into the model of an event or phenomenon
- hidden Markov model - HMM
- HMMs require building a profile based on a <u>training data set</u>
- Like other “machine learning” approaches.
- More Sensitive

HMM starts with a Multiple Sequence Alignment

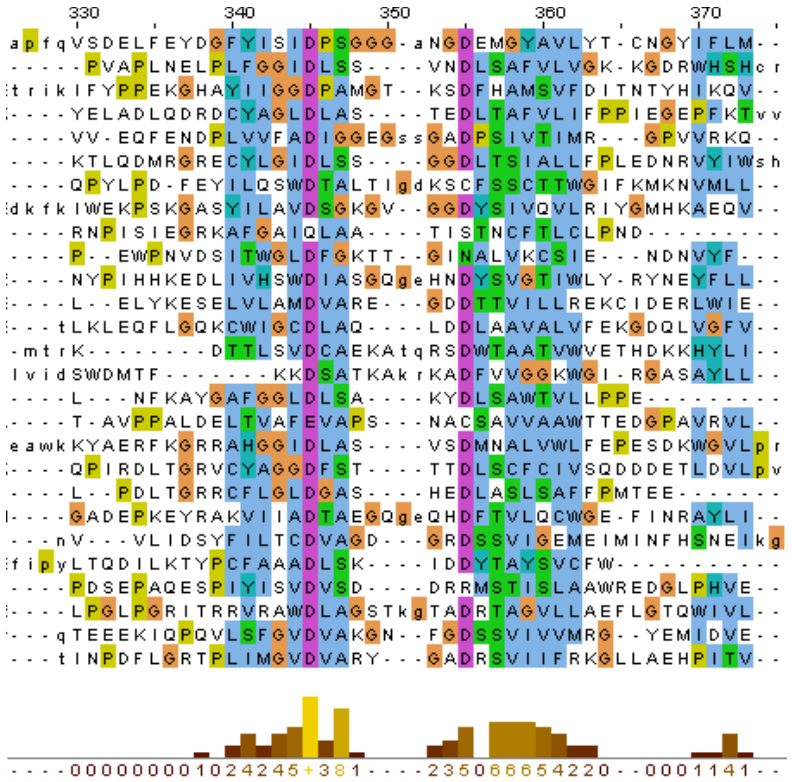
```
CLUSTAL 2.0.12 multiple sequence alignment

sp|088479|FOS_MESAU      MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLSVSSANF 60
sp|Q56TN0|FOS_PHORO      MMFSGFNTDYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCADLSVSSANF 60
sp|077628|FOS_BOVIN      MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDYCTDLAVSSANF 60
sp|Q8HZP6|FOS_FELCA      MMFSGFNADYEASSSRCSSASPAGDNLSYYHSPADSFSSMGSPVNAQDFCTDLAVSSANF 60
sp|P01100|FOS_HUMAN      MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNAQDFCTDLAVSSANF 60
sp|P12841|FOS_RAT        MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
sp|P01102|FOS_MSVFB      MMFSGFNADYEASSFRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
sp|P11939|FOS_CHICK      MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF 60
sp|P53539|FOSB_HUMAN     -MFQAFPGDYDSGS-RCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
sp|Q9TUB3|FOSB_CANFA     -MFQAFPGDYDSGS-RCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
sp|P13346|FOSB_MOUSE     -MFQAFPGDYDSGS-RCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
                          *.  *  .*.. * **** **. .    * *..***.*  :.. :*: *:.*.  ...*

sp|088479|FOS_MESAU      IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGVPTPS-----TGAYSR 108
sp|Q56TN0|FOS_PHORO      IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGVPTPS-----TGAYSR 108
sp|077628|FOS_BOVIN      IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGVPTPS-----AGAYSR 108
sp|Q8HZP6|FOS_FELCA      IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGVPAPS-----AGAYSR 108
sp|P01100|FOS_HUMAN      IPTVTAISTSPDLQWLVQPALVSSVAPS-------QTRAPHPFGVPAPS-----AGAYSR 108
sp|P12841|FOS_RAT        IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGLPTPS-----TGAYAR 108
sp|P01102|FOS_MSVFB      IPTVTATSTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGLPTQS-----AGAYAR 108
sp|P11939|FOS_CHICK      VPTVTAISTSPDLQWLVQPTLISSVAPS-------QNRG-HPYGVPAPAP----PAAYSR 108
sp|P53539|FOSB_HUMAN     VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSS 114
sp|Q9TUB3|FOSB_CANFA     VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGMSGYSS 114
sp|P13346|FOSB_MOUSE     VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYST 114
                         ;***** ;** ********;*;**;  *       *    .*;.;*  :      ..*:

sp|088479|FOS_MESAU      --------AGMVKTVSGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|Q56TN0|FOS_PHORO      --------AGMVKTVSGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|077628|FOS_BOVIN      --------AGVMKTMTGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|Q8HZP6|FOS_FELCA      --------AGVVKTVTAGG---RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 157
sp|P01100|FOS_HUMAN      --------AGVVKTMTGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|P12841|FOS_RAT        --------AGVVKTMSGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|P01102|FOS_MSVFB      --------AEMVKTVSGG----RAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
sp|P11939|FOS_CHICK      --------PAVLK-APGG----RGQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 155
sp|P53539|FOSB_HUMAN     GGASGSGGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
sp|Q9TUB3|FOSB_CANFA     GGASGSGGPSTSGTTSGPGPARPARARLRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
sp|P13346|FOSB_MOUSE     GGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
                                .      ..      .::  ** : * *:********:******:*******

sp|088479|FOS_MESAU      RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEM 216
sp|Q56TN0|FOS_PHORO      RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEDM 216
sp|077628|FOS_BOVIN      RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEM 216
sp|Q8HZP6|FOS_FELCA      RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEM 217
sp|P01100|FOS_HUMAN      RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEM 216
```



- Use structural and mechanistic information (e.g., catalytic sites)
- More sensitive than pairwise – detect distant relationships

Protein sequences can consist of structurally different parts

EXAMPLE

Domain

part of the tertiary structure of a protein that can exist, function and evolve independently of the rest, linked to a certain biological function

ATP-binding domain

Motif

part (not necessarily contiguous) of the primary structure of a protein that corresponds to the signature of a biological function. Can be associated with a domain.

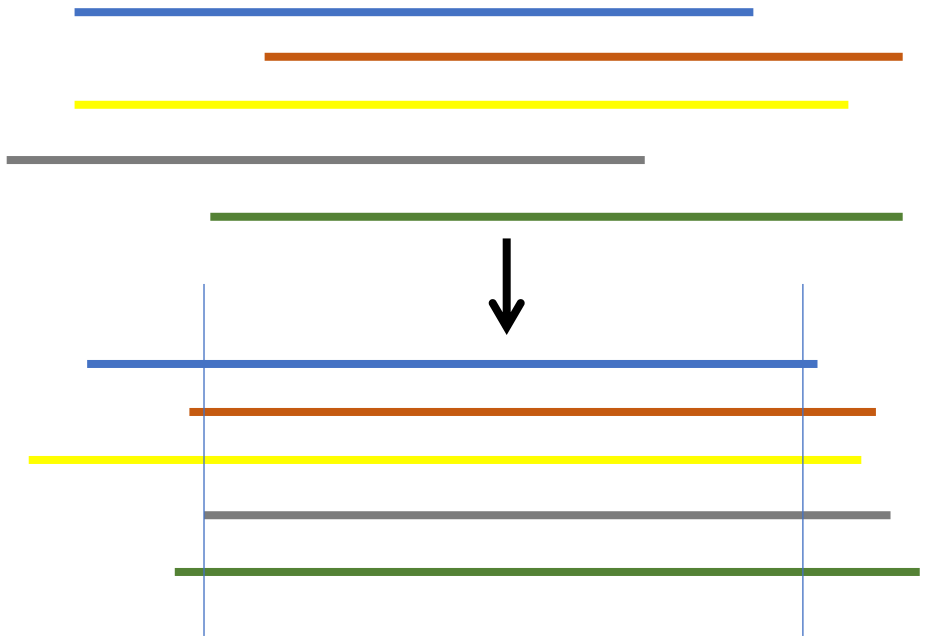
Heme motif CHXXH

Feature

part of the sequence for which some annotation has been added. Some features correspond to domain or motif assignments.

MSA to HMM or PSSM profile

Collect “seed” proteins



Generate & Trim Alignment

Region of good alignment and closest similarity

Generate Profile with HMM or PSSM

Compute statistical probabilities for amino acid patterns in the seed

Search New Model against all proteins

*Choose “noise” and **“trusted”** cutoff scores based on “known” versus “unknown” protein scores*

Frequency matrices or profiles include the chance of observing the residues

For every position of a motif, a list of all amino acids is made with their frequency. Position-specific weight/scoring matrix or profile. More sensitive way.

```
        123456
        ATPKAE
        KKPKAA
        AKPKAK
        TKPKPA
        AKPKT-
        AKPAAK
        KLPKAD
        AKPKAA
Consensus: AKPKA-
```

Profile

Position:	1.	2.	3.	4.	5.	6.
A	0.625	0	0	1/8	6/8	3/8
D	0	0	0	0	0	1/8
E	0	0	0	0	0	1/8
K	0.25	6/8	0	7/8	0	2/8
L	0	1/8	0	0	0	0
P	0	0	1	0	1/8	0
T	1/8	1/8	0	0	1/8	0
-	0	0	0	0	0	1/8
Sum	1	1	1	1	1	1

? Query: AKPKTE

? Query: KKPETE

? Query: TLPATE

Example: http://expasy.org/prosite/PS51092 http://prosite.expasy.org/prosuser.html#meth2

How good a sequence matches a profile is reported with a score

```
123456
ATPKAE
KKPKAA
AKPKAK
TKPKPA
AKPKT-
AKPAAK
KLPKAD
AKPKAA
```

Consensus: AKPKA-

PSWM: scores

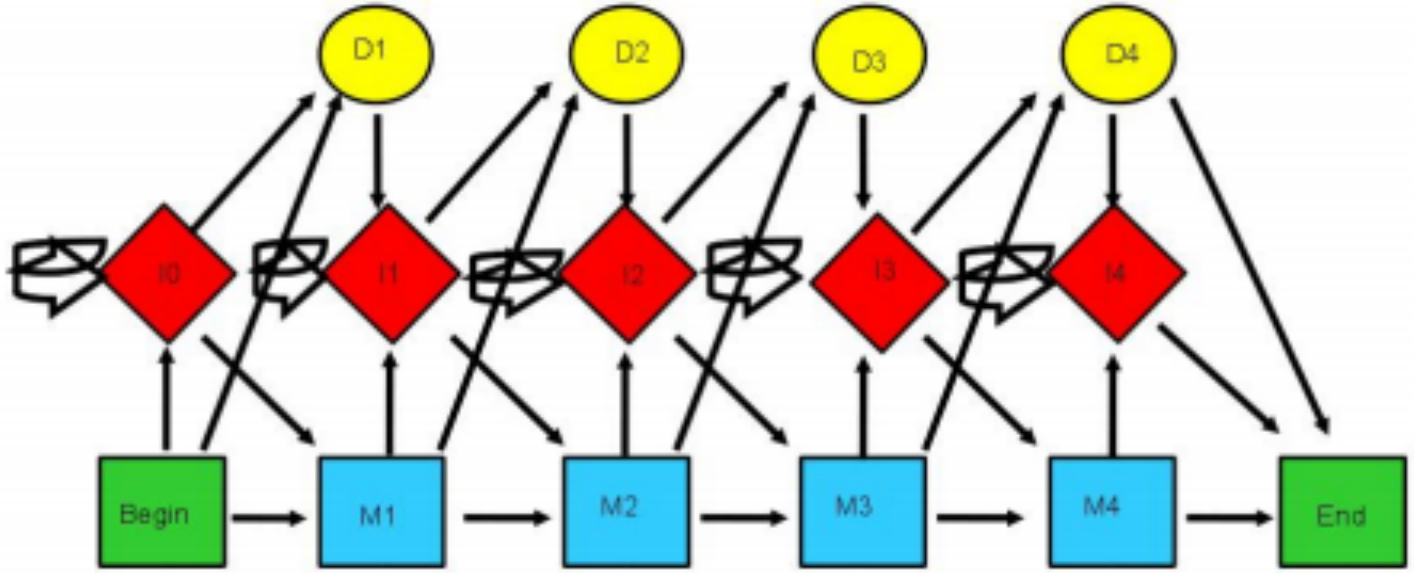
Position:	1.	2.	3.	4.	5.	6.
A	2.377	-2.358	-2.358	0.257	2.631	1.676
D	-2.358	-2.358	-2.358	-2.358	-2.358	0.257
E	-2.358	-2.358	-2.358	-2.358	-2.358	0.257
K	1.134	2.631	-2.358	2.847	-2.358	1.134
L	-2.358	0.257	-2.358	-2.358	-2.358	-2.358
P	-2.358	-2.358	0.257	-2.358	0.257	-2.358
T	0.257	0.257	-2.358	-2.358	0.257	-2.358

? Query: AKPKTE	Score = 11.4
? Query: KKPETE	Score = 5.0
? Query: TLPATE	Score = 4.3

http://prosite.expasy.org/prosuser.html#meth2

A hidden Markov Model takes also into account the gaps in an alignment

The schematic representation of a HMM

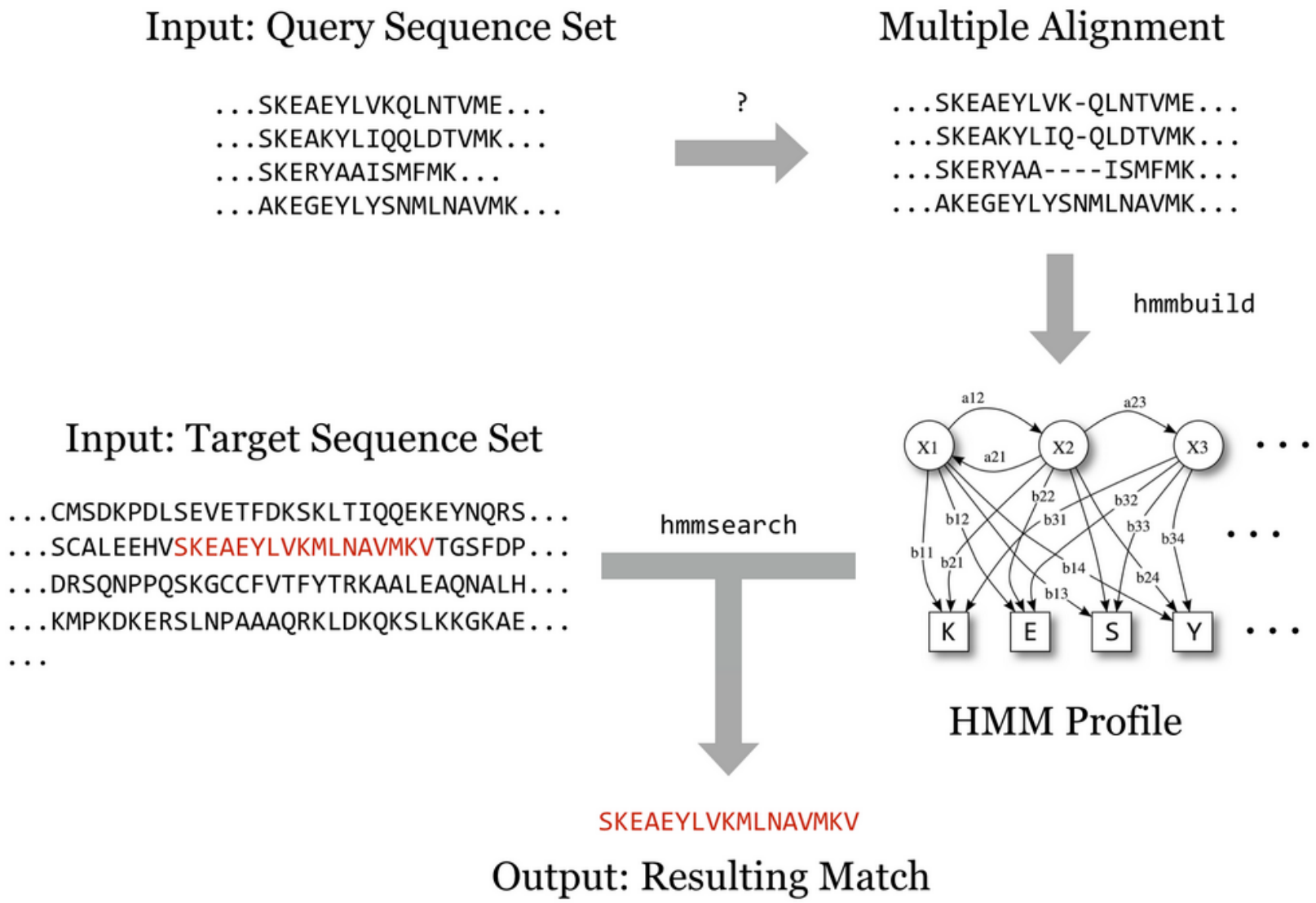


HMMER

biosequence analysis using profile hidden Markov models

http://www.myoops.org/twocw/mit/NR/rdonlyres/Electrical-Engineering

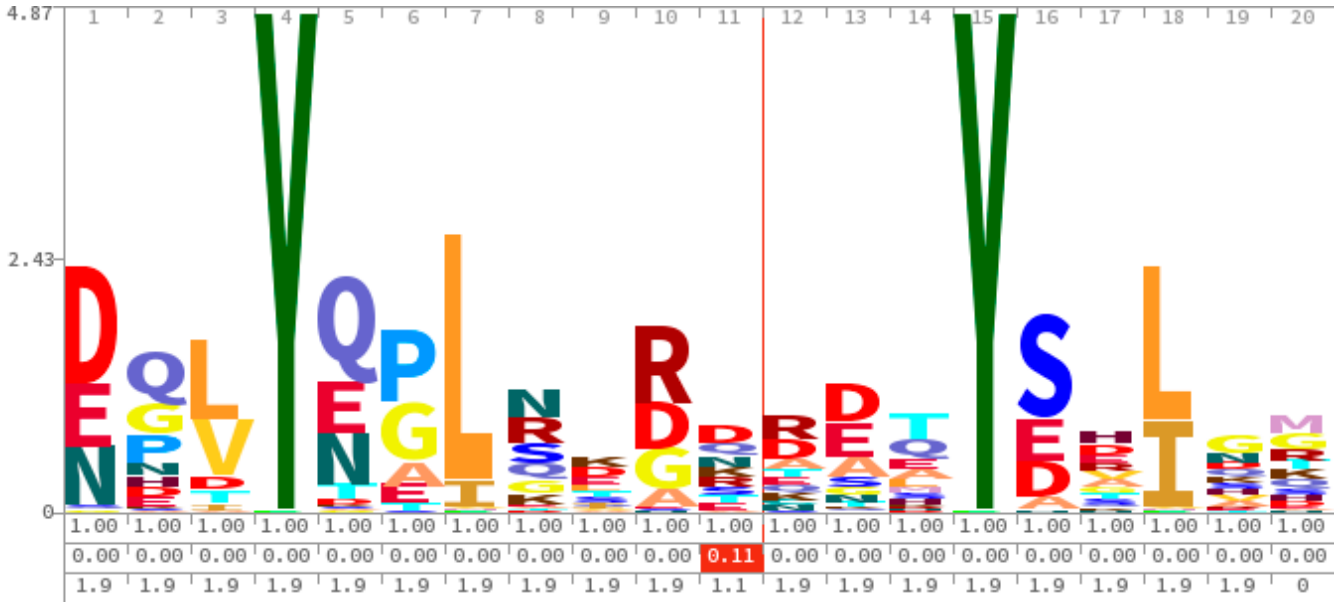
MSA to HMM profile using HMMER



Shawn O'Neil, CGRB

HMMER

HMMER is used for searching sequence databases for sequence homologs and for making sequence alignments.



Program	Task
hmmalign	build an MSA from sequences
hmmbuild	convert MSA to HMM
phmmer	single protein vs protein DB (like-BLAST)
hmmscan	protein sequence vs HMM database
hmmsearch	HMM database vs protein sequence
hmmconvert	convert HMM profile formats
hmmfetch	grab a single HMM from a database
hmmpress	compress HMM into binary format for hmmscan

Difference between hmmscan and hmmsearch

- hmmscan and hmmsearch are doing exactly the same compute
  - comparing one profile to one sequence at a time
  - bit score results are identical
  - save both in tabular output files
- hmmscan needs to read in all HMM profiles in a database for each query protein being searched → compute time << I/O time
  - hmmpress solves some of this issue
- hmmsearch reads in the HMM profile database so the search is now per query protein

PSI-BLAST & RPS-BLAST (NCBI)

PSI-BLAST

Position-Specific Iterated BLAST

- finds sequences significantly similar to the query in a database search
- Significant matches are used to make an alignment
- MSA used to build a Position-Specific Score Matrix (PSSM) for the query
- The new PSSM is searched against the same database again to pull in more significant hits based on conserved features
- Can be used to further refine the scoring model

RPS-BLAST

Reverse Position-Specific BLAST

- query sequence to search a database of pre-calculated PSSMs
  - reports significant hits in a single pass
- The PSSM has changed from "query" to "subject”
- Commonly used to search the NCBI CDD (Conserved Domain Database)

Lesson 4

Reference databases – series of lectures and examples about databases with a variety of functions

- **Pfam**
- **KEGG**
- **eggNOG**

Reference databases – lectures and examples about databases for specific functions

- **CAZy**
- **anitSMASH**