Biochem 3BP3

Sequence Similarity and Searching

Week of Sept 20, 2021

Why Sequence Analysis?

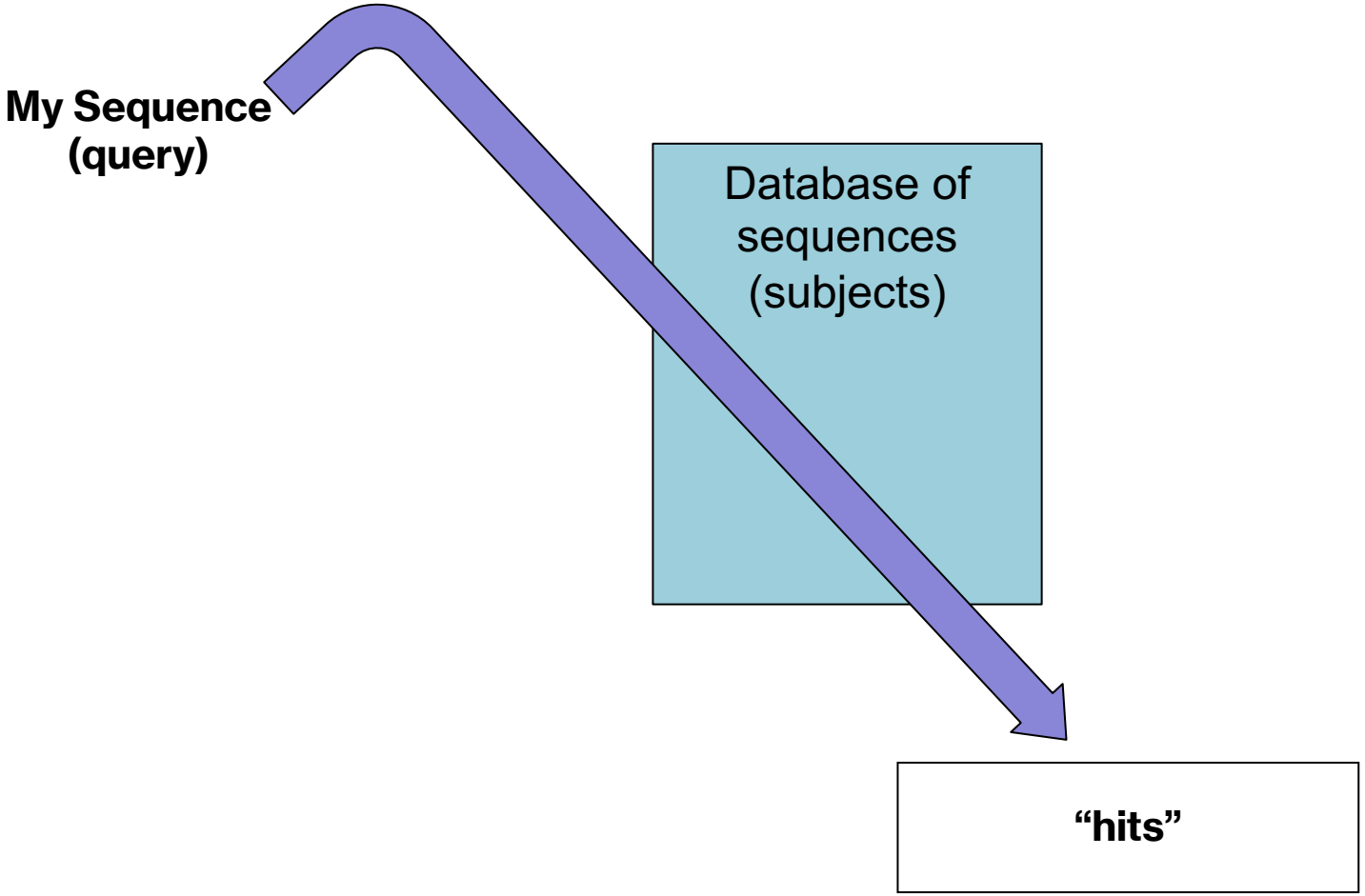
A. I have obtained a DNA sequence via PCR and Sanger sequencing – did I amplify the right sequence?
B. I have been sequencing a genome and have predicted Open Reading Frames and I want to know what they encode
C. I want to find my gene of interest in a genome sequence
D. I want to predict functional domains or motifs for my protein sequence
E. I want to know which regulatory binding sites are 5’ of my gene

There are many methods – we'll focus on three

- Local sequence alignment, e.g. BLAST
- Hidden Markov Models, e.g. Pfam/Hmmer
- Motif detection, e.g. PROSITE & PSSMs

Course Goal – understand how they work and how they differ

Basic Local Alignment Search Tool (BLAST)



Basic Local Alignment Search Tool (BLAST)

- BLAST is one of the workhorses of bioinformatics
- An approximation of the Smith-Waterman algorithm with an emphasis on efficiency and generalization
- Published in 1990
  - DNA sequence databases were coming online and growing in size
  - accessible computational power was a concern so a ‘fast’ algorithm was an important advance
- By 2000 fast computer chips and affordable parallel computing (i.e. many processors) made high-throughput BLAST very workable
- Today, advances in Next Generation Sequencing are exceeding Moore’s Law
  - BLAST is becoming slow again not because of the algorithm but because of the size of databases
  - This is an active time in new algorithm development (e.g. BLAT, DIAMOND)

Basic Local Alignment Search Tool (BLAST)

- A great deal of computer science and mathematics are inside BLAST
  - Scoring matrices
  - Search heuristics
  - Processor and memory usage
  - Database formatting and indexing
  - Data and File formats (INPUT and OUTPUT)

- Key Concepts
  - Searching for local alignment (versus global alignment)
    - Caveats for prediction of function
  - What is your question? “sequence space”
    - BLASTN, BLASTP, BLASTX, TBLASTN, TBLASTX
  - Similarity scoring – bitscore versus percent identity
    - Use of substitution matrices
  - Significance and the Expectation value (e-value)

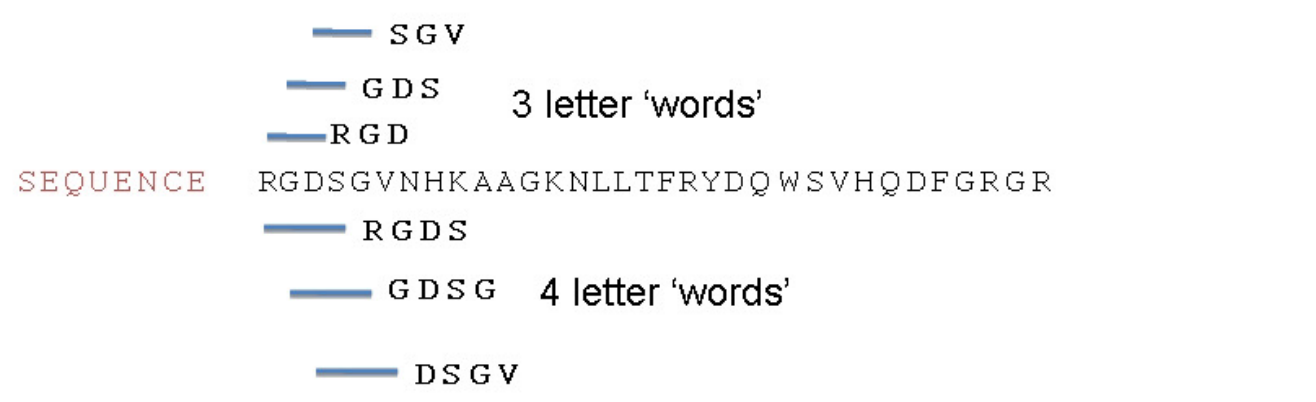
Before BLAST: Smith-Waterman

- dynamic programming alignment algorithms to compare the query against each sequence in the database
- each comparison is an exhaustive comparison of each nucleotide or amino acid against all others
- it won’t miss anything, but processing and memory intensive

Basic Local Alignment Search Tool (BLAST)

- novel decrease in the search space
- creates a “word list” from the query sequence with words of a specific length (*w*)
- local alignments only explored where “words” have complete match to the query
- short word matching is very amenable to computing – fast and lower memory needs

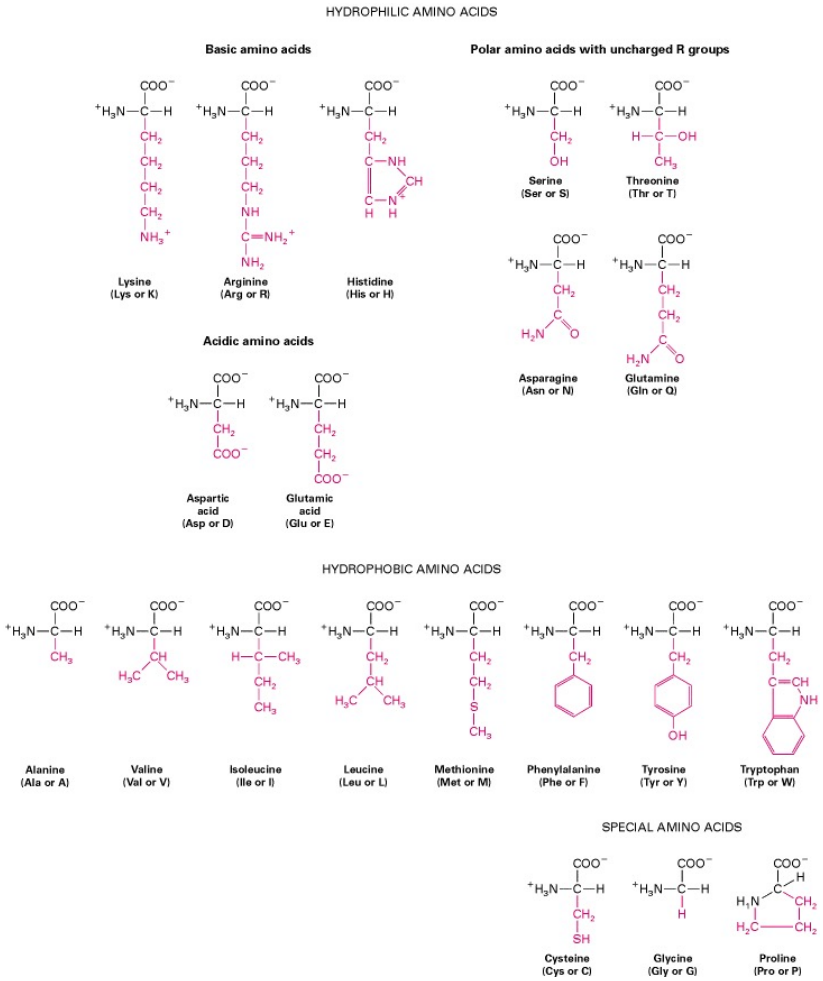
Basic Local Alignment Search Tool (BLAST)



- smaller word sizes provide better resolution but there are more of them so they increase analysis time.
- BLAST defaults (*w*=11 for DNA, *w*=3 for protein) are often sufficient – but not always!

How does BLAST determine if there is match?

It uses a cut-off score based on a scoring matrix



conservative = similar physico-chemical properties

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

- BLOSUM62 matrix (BLOcks SUbstitution Matrix) reflects the relative rate of substitution among all 20 amino acids observed in conserved regions (no more than 62% similarity) of known protein sequences.
- BLOSUM62 is the BLAST default. Since is it based on conserved regions with 62% similarity **or less** it is among the best for detecting most weak protein similarities.
- Other BLOSUM or PAM matrices exist for detection of less or more divergent proteins.

BLAST use of scoring matrix

Abbreviation	1 letter abbreviation	Amino acid name
Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Pyl	O	Pyrrolysine
Ser	S	Serine
Sec	U	Selenocysteine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine
Asx	B	Aspartic acid or Asparagine
Glx	Z	Glutamic acid or Glutamine
Xaa	X	Any amino acid
Xle	J	Leucine or Isoleucine
TERM		termination codon

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Query NYLENFVQATFN

Query words NYL YLE LEN ENF NFV FVQ VQA QAT ATF TFN

```
Query          ENF
Subject   SSTNYAENTIQSIISTVEPAQR
```

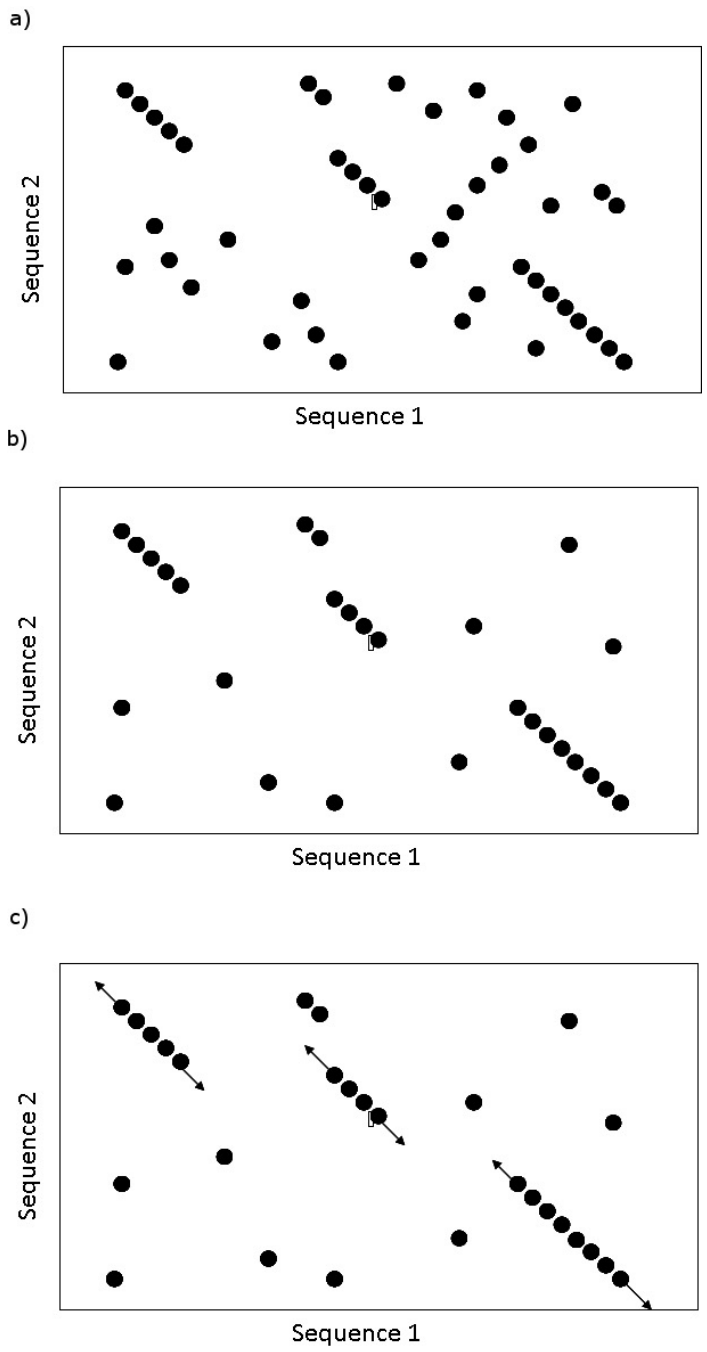
Seed

```
Query     NLYENFVQATFNALTAEKV
          NY ENF+Q+   + +    +
Subject   NYAENTIQSIISTVEPAQR
```

Alignment extended as long as the score doesn't go below the cut-off. Called a High Scoring Pair (HSP)

Basic Local Alignment Search Tool (BLAST)

- the seed alignment (*w*) is then extended based on extension / scoring criteria – defaults are often used
- extension is tolerant of gaps
- by using seed alignments, BLAST solves local alignments within the query/subject pair – not alignment along the entire sequence
- local alignments are called a High Scoring Pair (HSP). The example at the right has three HSPs.



A High Scoring Pair (HSP)

DNA-binding response regulator [Burkholderia cenocepacia]

Sequence ID: ref|WP_050014536.1| Length: 220 Number of Matches: 1

Range 1: 2 to 214 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
184 bits(466)	7e-54	Compositional matrix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

```
Query  3    KILMIEDDFKIAESTITLLQYHQFEVEWVNNGLDGLAQLAKTKFDLILLDLGLPMMDGMQ  62
            +IL++EDD  IAE     L+   F V+WV +G   L  L    +DL+LLDLGLP  DG+
Sbjct  2    RILLVEDDRMIAEGVRKALRSDGFAVDWVQDGDAALTALGGETYDLLLLDLGLPKRDGID  61

Query  63   VLKQIRQRA-ATPVLIISARDQLQNRVDGLNLGADDYLIKPYEFDELLARIHALLRRSGV  121
            VL+ +R R  A PVLI++ARD + +RV GL+ GADDYL+KP++ DEL AR+ AL+RR
Sbjct  62   VLRTLRGRGLALPVLIVTARDAVADRVKGLDAGADDYLVKPFDLDELGARMRALIRR---  118

Query  122  EAQLASQDQLLESGDLVLNVEQHIATFKGQRIDLSNREWAILIPLMTHPNKIFSKANLED  181
              Q    + L+  G L L+   H  T  G  + LS RE+A+L  L+  P  + SK+ LE+
Sbjct  119  --QAGRSESLIRHGALTLDPAAHQVTLDGAPVALSAREFALLEALLARPGAVLSKSQLEE  176

Query  182  KLYDFDSDVTSNTIEVYVHHLRAKLGKDFIRTIRGLGY  219
            K+Y +  ++ SNT+EVY+H LR KLG D IR +RGLGY
Sbjct  177  KMYGWGEEIGSNTVEVYIHALRKKLGSDLIRNVRGLGY  214
```


Score	Expect	Method	Identities	Positives	Gaps
184 bits(466)	7e-54	Compositional matrix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

219 – (2) + 1 gap = 218

3rd amino acid is a part of the alignment

A High Scoring Pair (HSP)

DNA-binding response regulator [Burkholderia cenocepacia]

Sequence ID: ref|WP_050014536.1| Length: 220 Number of Matches: 1

Range 1: 2 to 214 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
		ix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

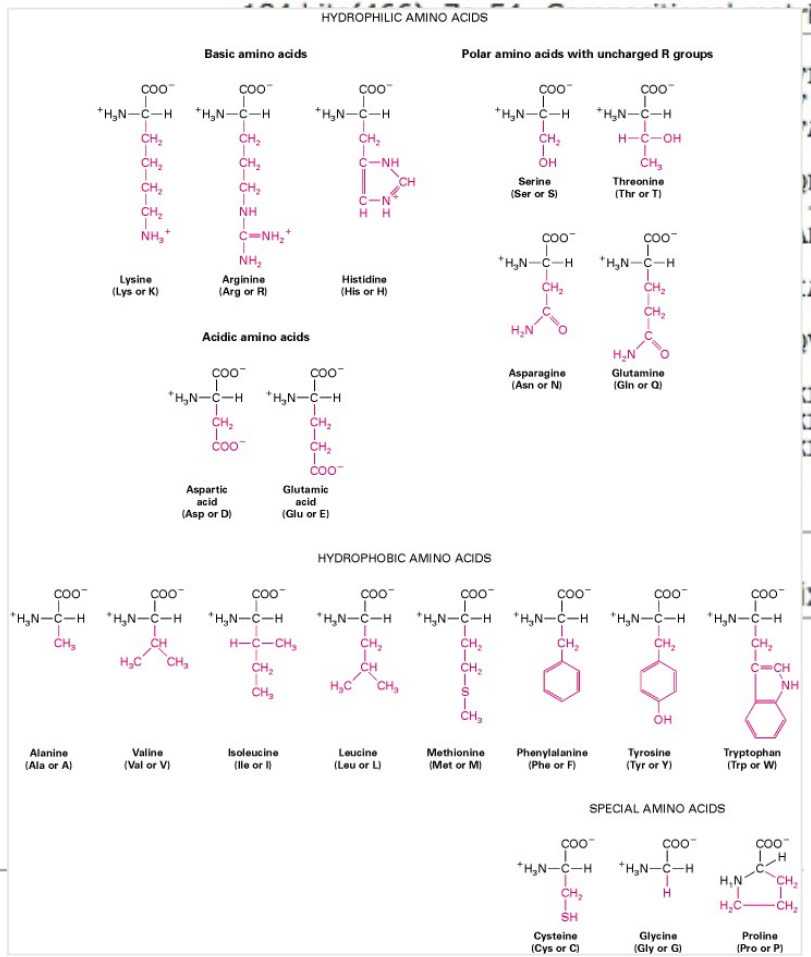
```
EVEWVNNGLDGLAQLAKTKFDLILLDLGLPMMDGMQ   62
 V+WV +G   L  L    +DL+LLDLGLP  DG+
AVDWVQDGDAALTALGGETYDLLLLDLGLPKRDGID   61

NRVDGLNLGADDYLIKPYEFDELLARIHALLRRSGV   121
 +RV GL+ GADDYL+KP++ DEL AR+ AL+RR
DRVKGLDAGADDYLVKPFDLDELGARMRALIRR---   118

ATFKGQRIDLSNREWAILIPLMTHPNKIFSKANLED   181
 T  G  + LS RE+A+L  L+  P  + SK+ LE+
VTLDGAPVALSAREFALLEALLARPGAVLSKSQLEE   176

LGKDFIRTIRGLGY   219
LG D IR +RGLGY
LGSDLIRNVRGLGY   214
```


	Identities	Positives	Gaps
ix adjust.	102/218(47%)	140/218(64%)	6/218(2%)



conservative = similar physico-chemical properties

conservative change

A High Scoring Pair (HSP)

DNA-binding response regulator [Burkholderia cenocepacia]

Sequence ID: ref|WP_050014536.1| Length: 220 Number of Matches: 1

Range 1: 2 to 214 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
		ix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

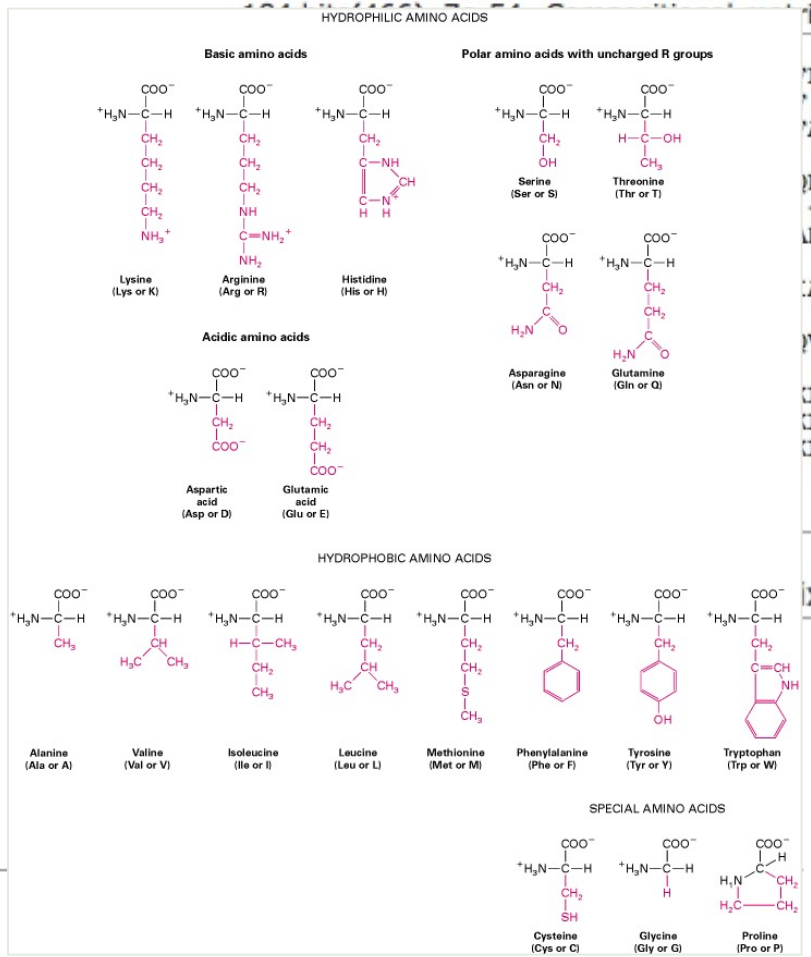
```
EVEWVNNGLDGLAQLAKTKFDLILLDLGLPMMDGMQ   62
 V+WV +G   L  L    +DL+LLDLGLP  DG+
AVDWVQDGDAALTALGGETYDLLLLDLGLPKRDGID   61

NRVDGLNLGADDYLIKPYEFDELLARIHALLRRSGV   121
 +RV GL+ GADDYL+KP++ DEL AR+ AL+RR
DRVKGLDAGADDYLVKPFDLDELGARMRALIRR---   118

ATFKGQRIDLSNREWAILIPLMTHPNKIFSKANLED   181
 T  G  + LS RE+A+L  L+  P  + SK+ LE+
VTLDGAPVALSAREFALLEALLARPGAVLSKSQLEE   176

LGKDFIRTIRGLGY   219
LG D IR +RGLGY
LGSDLIRNVRGLGY   214
```


	Identities	Positives	Gaps
ix adjust.	102/218(47%)	140/218(64%)	6/218(2%)



conservative = similar physico-chemical properties

exact matches

exact matches + conservative changes

gaps

A High Scoring Pair (HSP)

DNA-binding response regulator [Burkholderia cenocepacia]

Sequence ID: ref|WP_050014536.1| Length: 220 Number of Matches: 1

Range 1: 2 to 214 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
184 bits(466)	7e-54	Compositional matrix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

```
Query  3    KILMIEDDFKIAESTITLLQYHQFEVEWVNNGLDGLAQLAKTKFDLILLDLGLPMMDGMQ  62
            +IL++EDD  IAE     L+   F V+WV +G   L  L    +DL+LLDLGLP  DG+
Sbjct  2    RILLVEDDRMIAEGVRKALRSDGFAVDWVQDGDAALTALGGETYDLLLLDLGLPKRDGID  61

Query  63   VLKQIRQRA-ATPVLIISARDQLQNRVDGLNLGADDYLIKPYEFDELLARIHALLRRSGV  121
            VL+ +R R  A PVLI++ARD + +RV GL+ GADDYL+KP++ DEL AR+ AL+RR
Sbjct  62   VLRTLRGRGLALPVLIVTARDAVADRVKGLDAGADDYLVKPFDLDELGARMRALIRR---  118

Query  122  EAQLASQDQLLESGDLVLNVEQHIATFKGQRIDLSNREWAILIPLMTHPNKIFSKANLED  181
              Q    + L+  G L L+   H  T  G  + LS RE+A+L  L+  P  + SK+ LE+
Sbjct  119  --QAGRSESLIRHGALTLDPAAHQVTLDGAPVALSAREFALLEALLARPGAVLSKSQLEE  176

Query  182  KLYDFDSDVTSNTIEVYVHHLRAKLGKDFIRTIRGLGY  219
            K+Y +  ++ SNT+EVY+H LR KLG D IR +RGLGY
Sbjct  177  KMYGWGEEIGSNTVEVYIHALRKKLGSDLIRNVRGLGY  214
```


Score	Expect	Method	Identities	Positives	Gaps
184 bits(466)	7e-54	Compositional matrix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

The bit score gives an indication of how good the alignment is; a bit is a measure of information content; **the higher the score, the better the alignment**. Bit score uses identity, positives, and gaps (i.e. all data).

Bit score is independent of query sequence length and database size (i.e. normalized) allowing comparison among different searches or databases.

A High Scoring Pair (HSP)

DNA-binding response regulator [Burkholderia cenocepacia]

Sequence ID: ref|WP_050014536.1| Length: 220 Number of Matches: 1

Range 1: 2 to 214 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
184 bits(466)	7e-54	Compositional matrix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

```
Query  3    KILMIEDDFKIAESTITLLQYHQFEVEWVNNGLDGLAQLAKTKFDLILLDLGLPMMDGMQ  62
            +IL++EDD  IAE     L+   F V+WV +G   L  L    +DL+LLDLGLP  DG+
Sbjct  2    RILLVEDDRMIAEGVRKALRSDGFAVDWVQDGDAALTALGGETYDLLLLDLGLPKRDGID  61

Query  63   VLKQIRQRA-ATPVLIISARDQLQNRVDGLNLGADDYLIKPYEFDELLARIHALLRRSGV  121
            VL+ +R R  A PVLI++ARD + +RV GL+ GADDYL+KP++ DEL AR+ AL+RR
Sbjct  62   VLRTLRGRGLALPVLIVTARDAVADRVKGLDAGADDYLVKPFDLDELGARMRALIRR---  118

Query  122  EAQLASQDQLLESGDLVLNVEQHIATFKGQRIDLSNREWAILIPLMTHPNKIFSKANLED  181
              Q    + L+  G L L+   H  T  G  + LS RE+A+L  L+  P  + SK+ LE+
Sbjct  119  --QAGRSESLIRHGALTLDPAAHQVTLDGAPVALSAREFALLEALLARPGAVLSKSQLEE  176

Query  182  KLYDFDSDVTSNTIEVYVHHLRAKLGKDFIRTIRGLGY  219
            K+Y +  ++ SNT+EVY+H LR KLG D IR +RGLGY
Sbjct  177  KMYGWGEEIGSNTVEVYIHALRKKLGSDLIRNVRGLGY  214
```


Score	Expect	Method	Identities	Positives	Gaps
184 bits(466)	7e-54	Compositional matrix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

The expectation value (e-value) estimates the likelihood that a given sequence match is purely by chance. The lower the expectation value, the less likely the database match is a result of random chance and therefore the more significant.

E-value is a function of database size – how good is the database's sample of "sequence space" to determine random matches?

A High Scoring Pair (HSP)

DNA-binding response regulator [Burkholderia cenocepacia]

Sequence ID: ref|WP_050014536.1| Length: 220 Number of Matches: 1

Range 1: 2 to 214 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
184 bits(466)	7e-54	Compositional matrix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

```
Query  3    KILMIEDDFKIAESTITLLQYHQFEVEWVNNGLDGLAQLAKTKFDLILLDLGLPMMDGMQ  62
            +IL++EDD  IAE     L+   F V+WV +G   L  L    +DL+LLDLGLP  DG+
Sbjct  2    RILLVEDDRMIAEGVRKALRSDGFAVDWVQDGDAALTALGGETYDLLLLDLGLPKRDGID  61

Query  63   VLKQIRQRA-ATPVLIISARDQLQNRVDGLNLGADDYLIKPYEFDELLARIHALLRRSGV  121
            VL+ +R R  A PVLI++ARD + +RV GL+ GADDYL+KP++ DEL AR+ AL+RR
Sbjct  62   VLRTLRGRGLALPVLIVTARDAVADRVKGLDAGADDYLVKPFDLDELGARMRALIRR---  118

Query  122  EAQLASQDQLLESGDLVLNVEQHIATFKGQRIDLSNREWAILIPLMTHPNKIFSKANLED  181
              Q    + L+  G L L+   H  T  G  + LS RE+A+L  L+  P  + SK+ LE+
Sbjct  119  --QAGRSESLIRHGALTLDPAAHQVTLDGAPVALSAREFALLEALLARPGAVLSKSQLEE  176

Query  182  KLYDFDSDVTSNTIEVYVHHLRAKLGKDFIRTIRGLGY  219
            K+Y +  ++ SNT+EVY+H LR KLG D IR +RGLGY
Sbjct  177  KMYGWGEEIGSNTVEVYIHALRKKLGSDLIRNVRGLGY  214
```

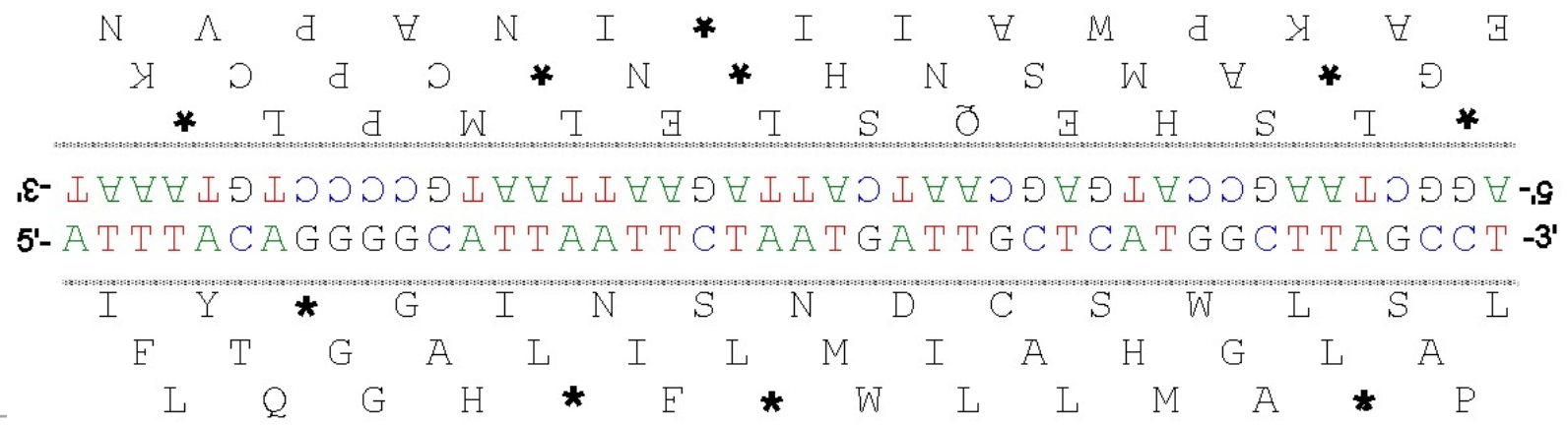

Score	Expect	Method	Identities	Positives	Gaps
184 bits(466)	7e-54	Compositional matrix adjust.	102/218(47%)	140/218(64%)	6/218(2%)

As a database grows, the same search will produce an altered expectation value. Different sized databases will produce different expectation values for the same HSPs.

There is no “best” expectation value but some generalizations are used: e^{-10} or smaller is worth examining; 0.01 or larger is noise.

BLAST Programs

- BLASTN – search a nucleotide database with a nucleotide query to find nucleotide HSPs
- BLASTP – search a protein database with a protein query to find protein HSPs
- BLASTX – search a protein database with a nucleotide query to find protein HSPs (translate the query in all six reading frames)



BLAST Programs

- BLASTN – search a nucleotide database with a nucleotide query to find nucleotide HSPs
- BLASTP – search a protein database with a protein query to find protein HSPs
- BLASTX – search a protein database with a nucleotide query to find protein HSPs (translate the query in all six reading frames)
- TBLASTN – search a nucleotide database with a protein query to find protein HSPs (translate the database in all six reading frames)
- TBLASTX – search a nucleotide database with a nucleotide query to find protein HSPs (translate the database & the query in all six reading frames)

BLAST is not Functional Biology

- A local alignment (HSP) found by BLAST may have little to do with protein function

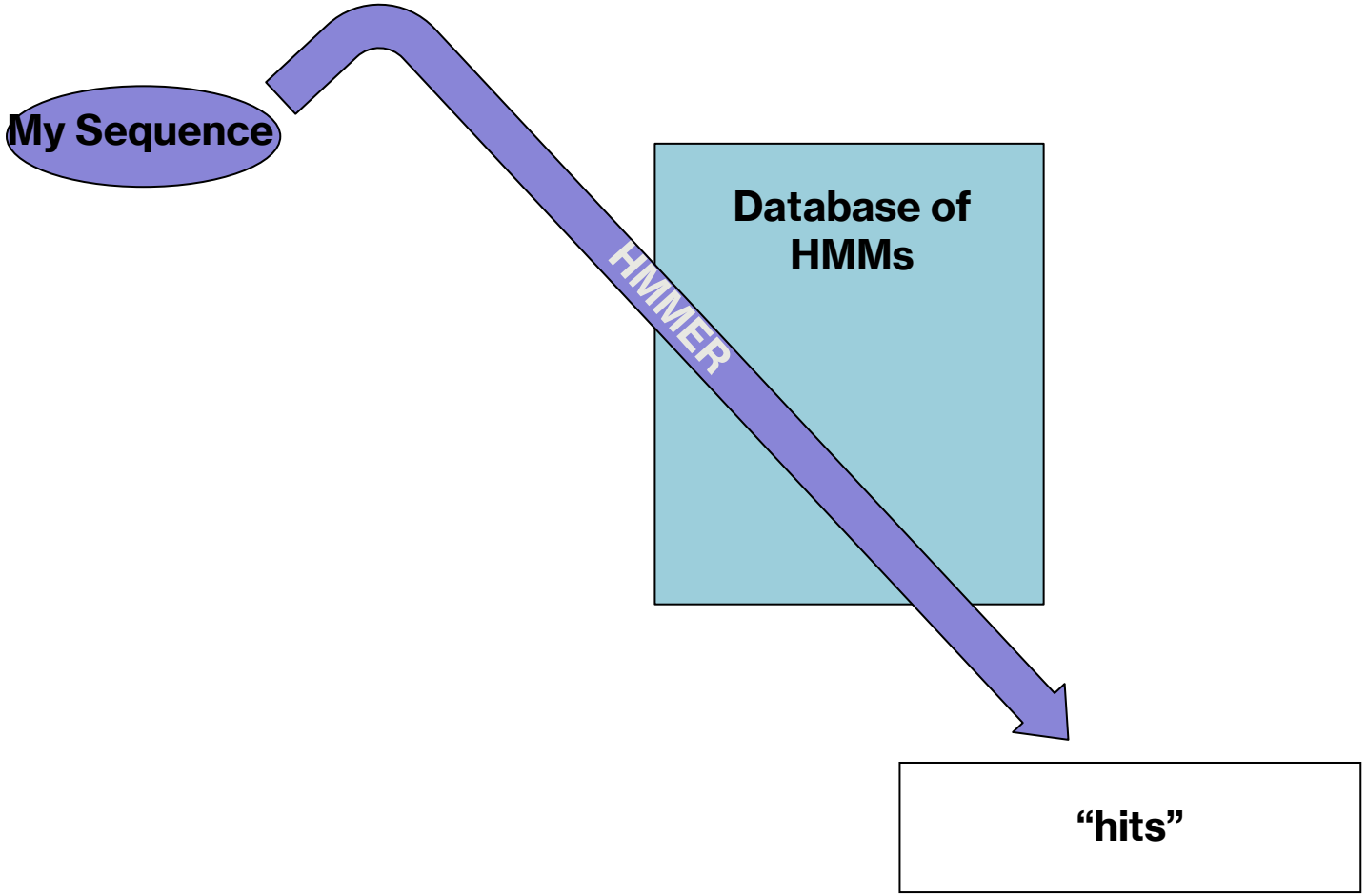
- BLAST knows about nucleotides, amino acids, and gaps but does not understand functional domains; it will not even detect functional domain similarity if it is outside of BLOSUM62 range

- Multi-domain proteins can give mis-leading BLAST results:

  - an ANT(3")-AAC(6') fusion protein will have BLAST hits to three types of proteins:
    - other ANT(3")-AAC(6') proteins
    - ANT(3") proteins
    - AAC(6') proteins

  - If the AAC(6') domain is poorly conserved, the query ANT(3")-AAC(6') fusion protein will only have good HSPs to:
    - ANT(3") proteins

Hidden Markov Models (HMMs)



Hidden Markov Models (HMMs)

- HMMs are not DNA or protein sequences but are models of how specific DNA or protein sequences are known to vary
  - e.g. an HMM for a iron hydrogenase domain

- A “hit” means your query sequence has an adequate fit to that model of variation

- Models are trained using real data, e.g. a sample of hydrogenase sequences

- Markov Models are probabilistic
  - every query has a probability of “fit” to the model
  - the probability is a function of a linear series of ‘labeling problems’

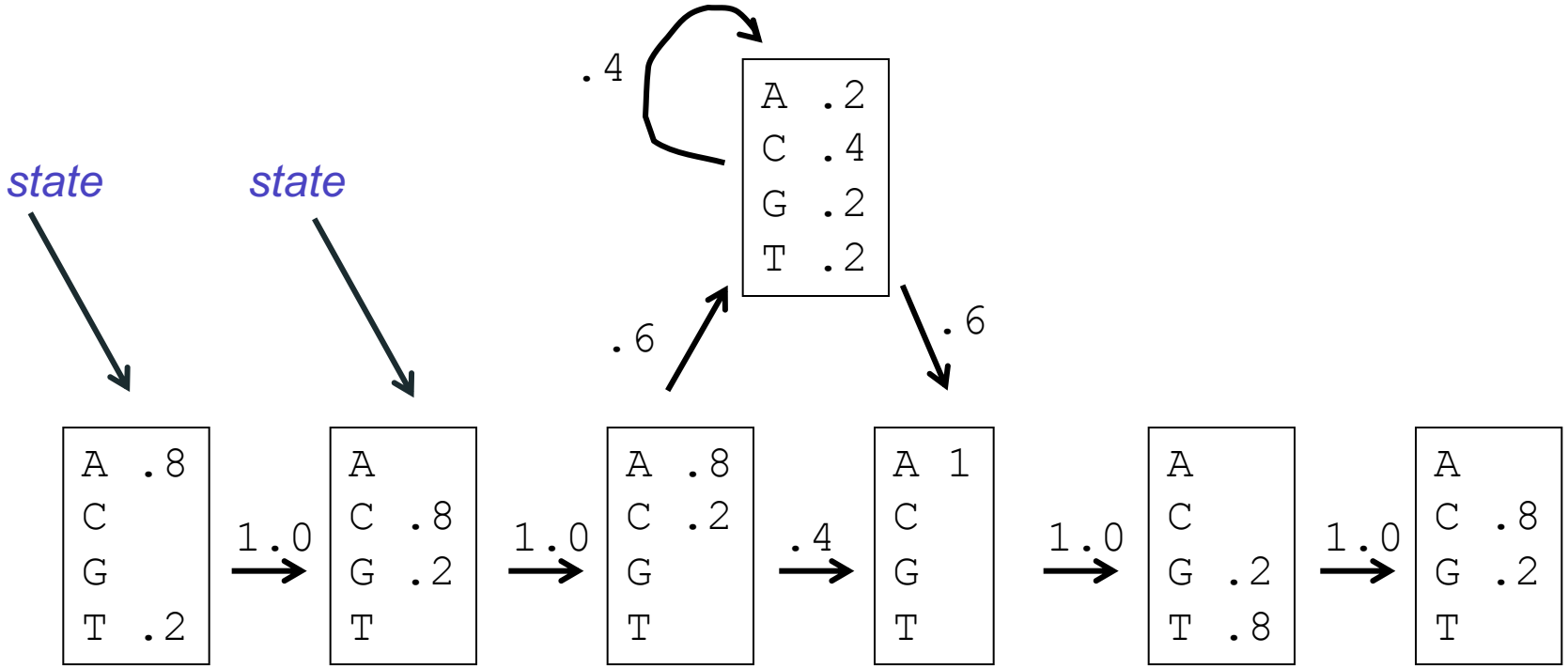
- Sequence HMMs focus on states with:
  - Emission probabilities (nucleotide / amino acid)
  - Transition probabilities (another nucleotide / amino acid or a gap)

A simple DNA HMM

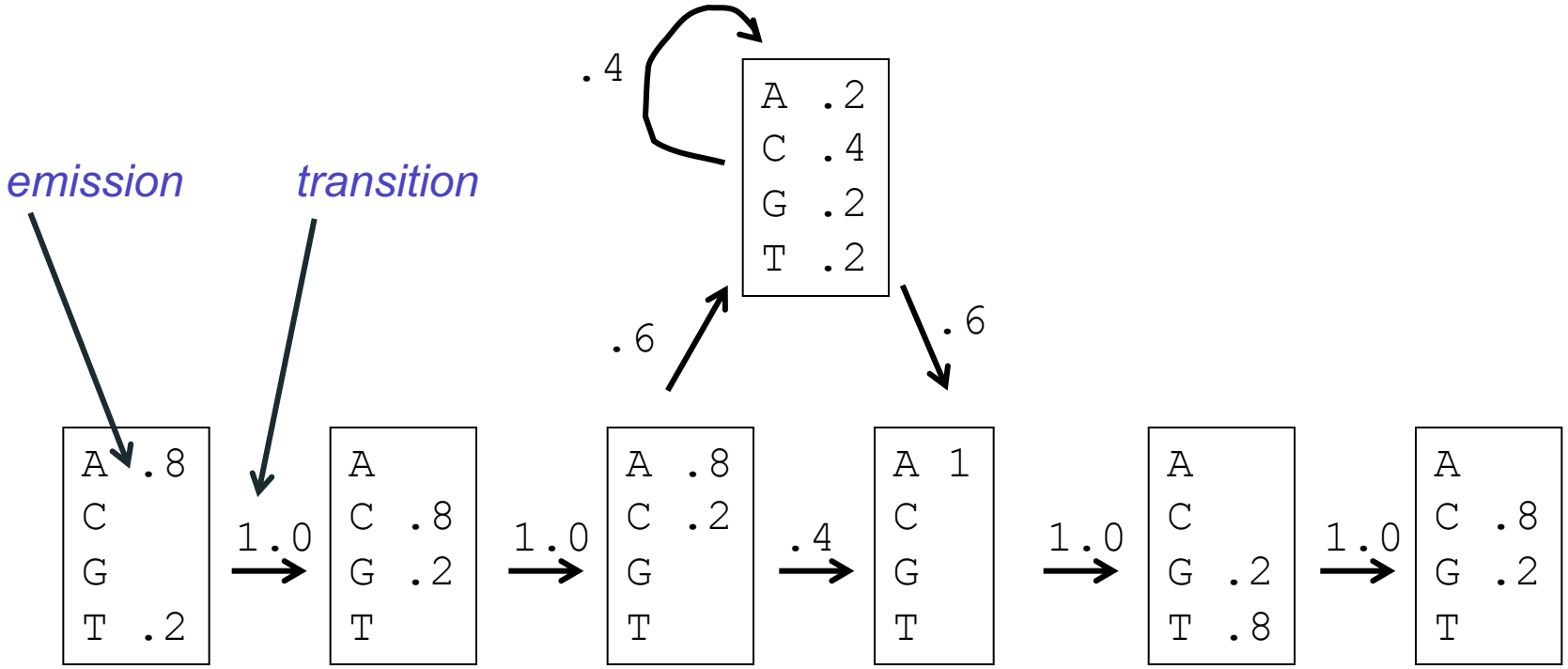
```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

5 species have slightly different DNA binding sites for a regulatory protein

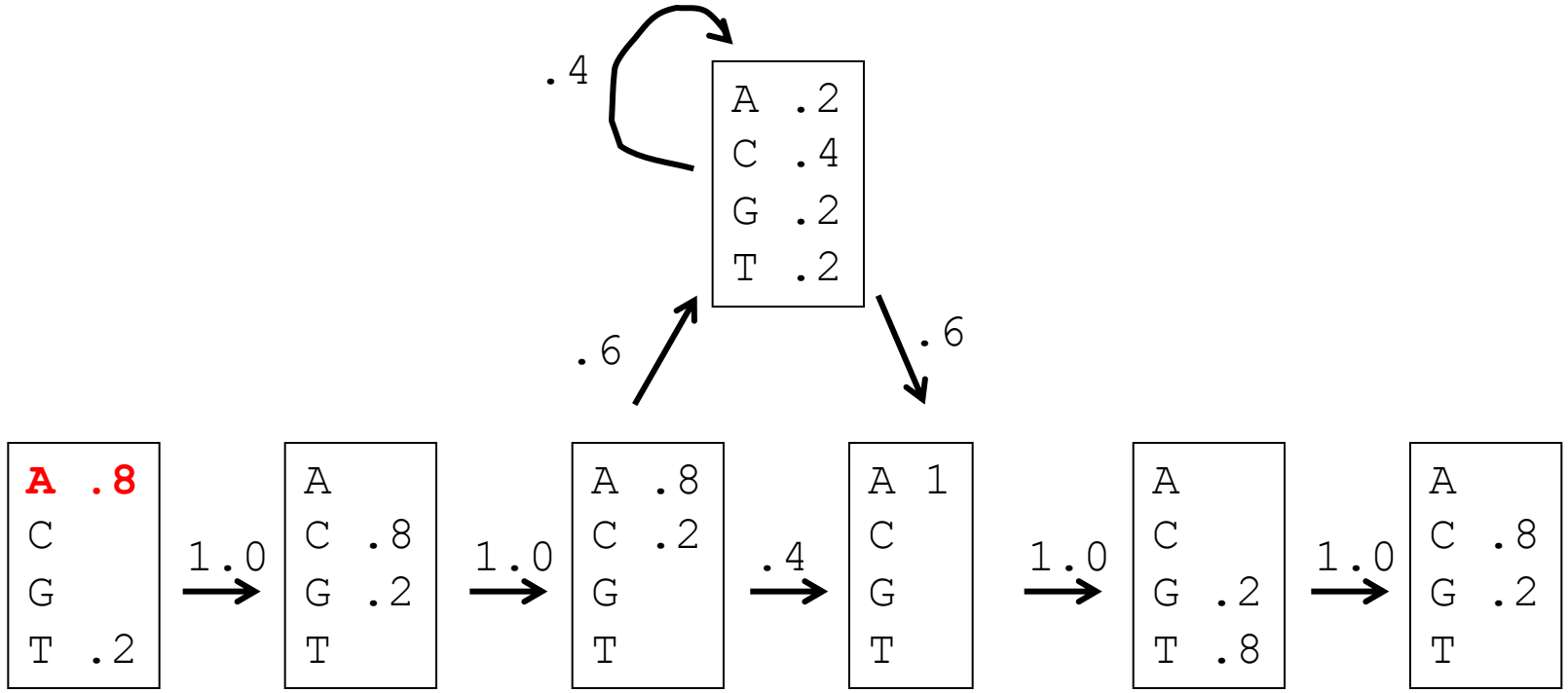
```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



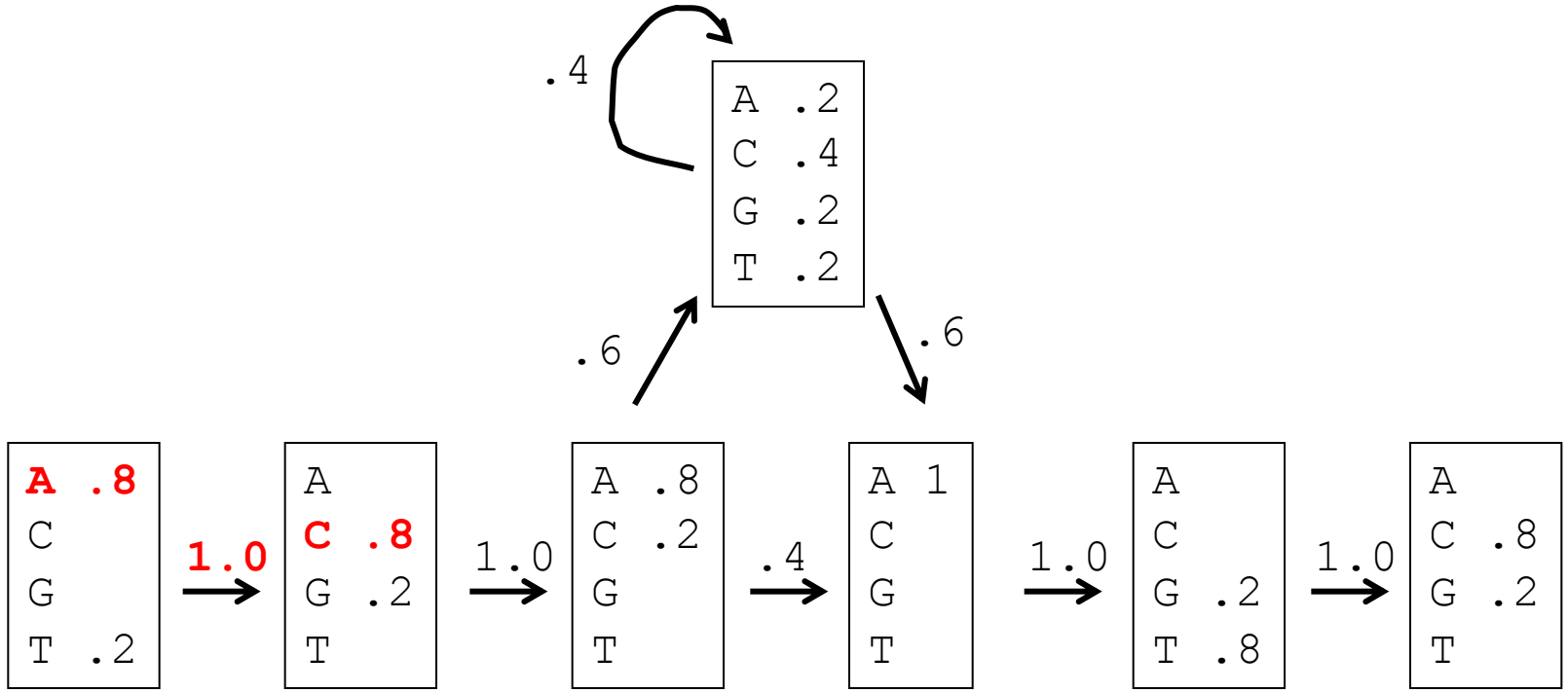
```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



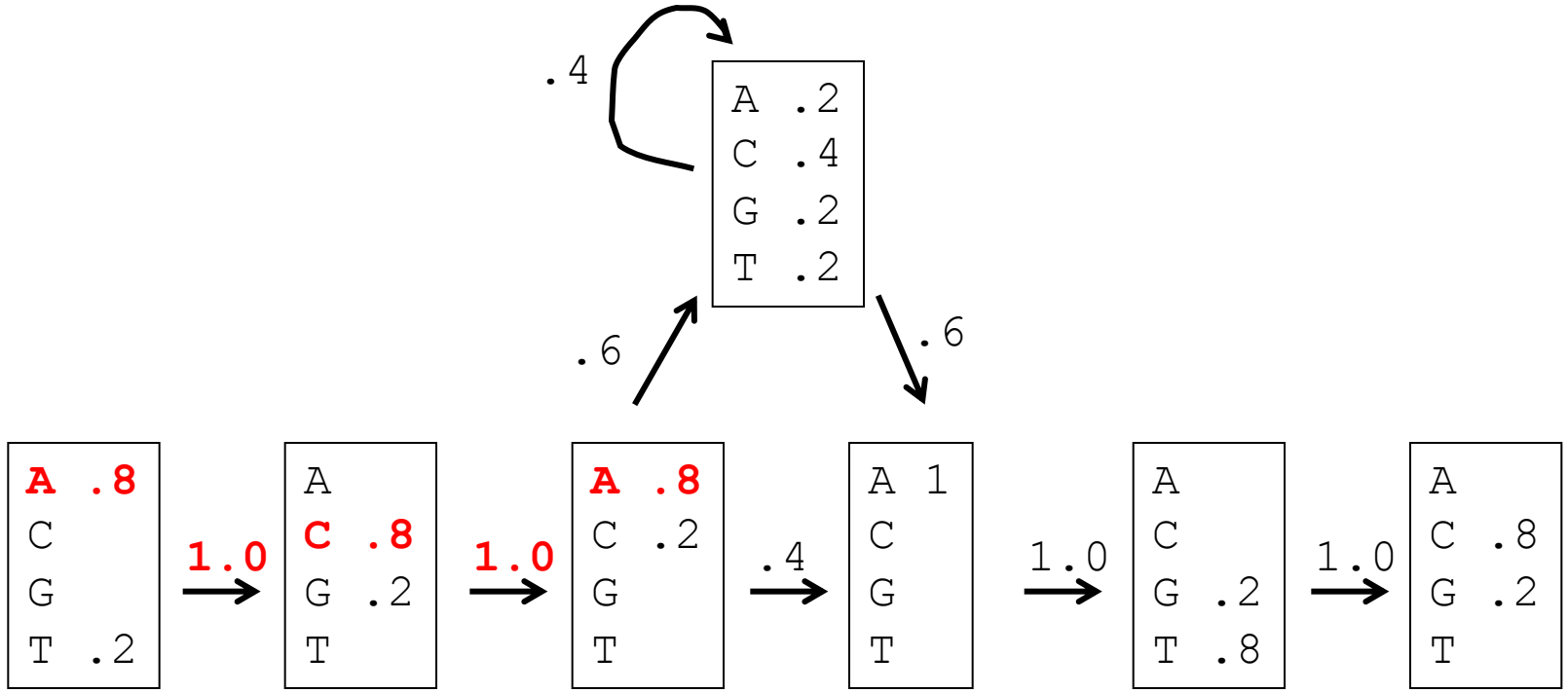
```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



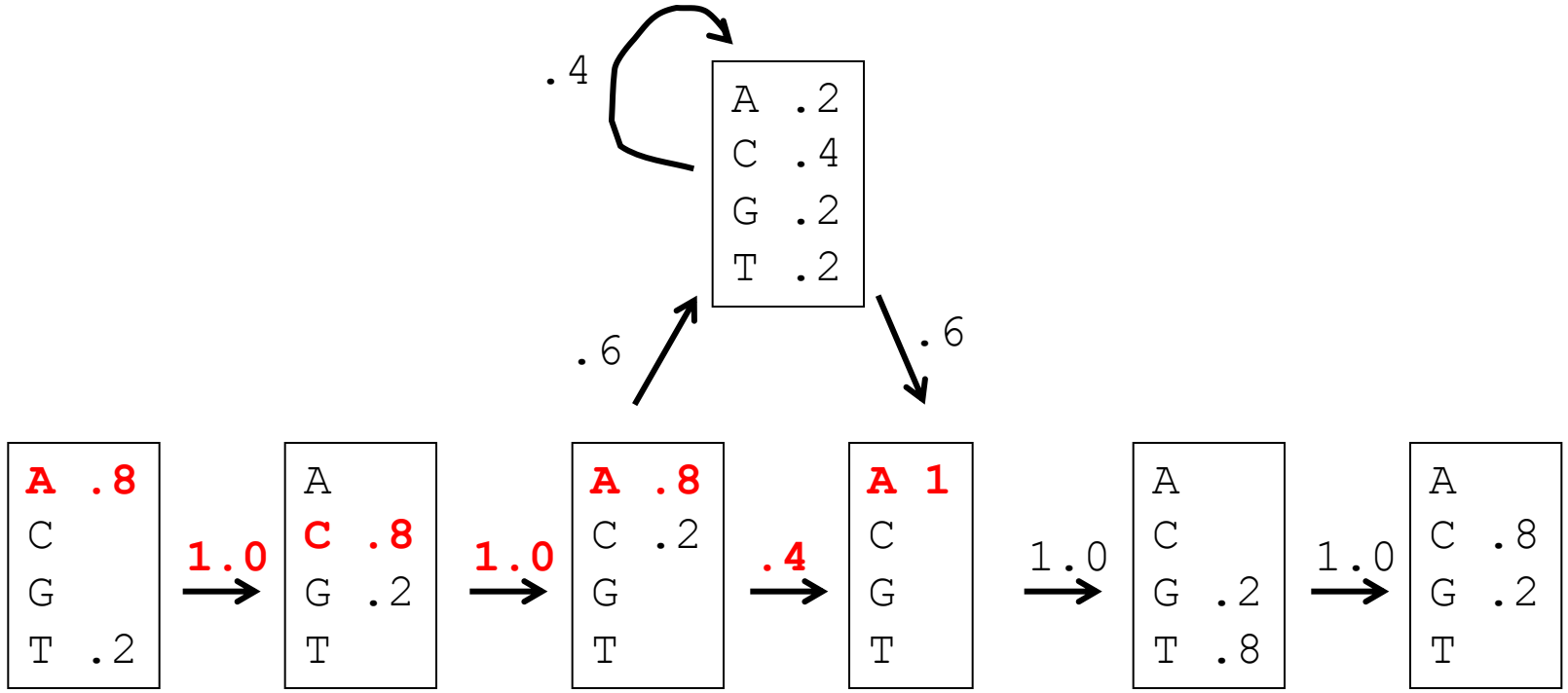
```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



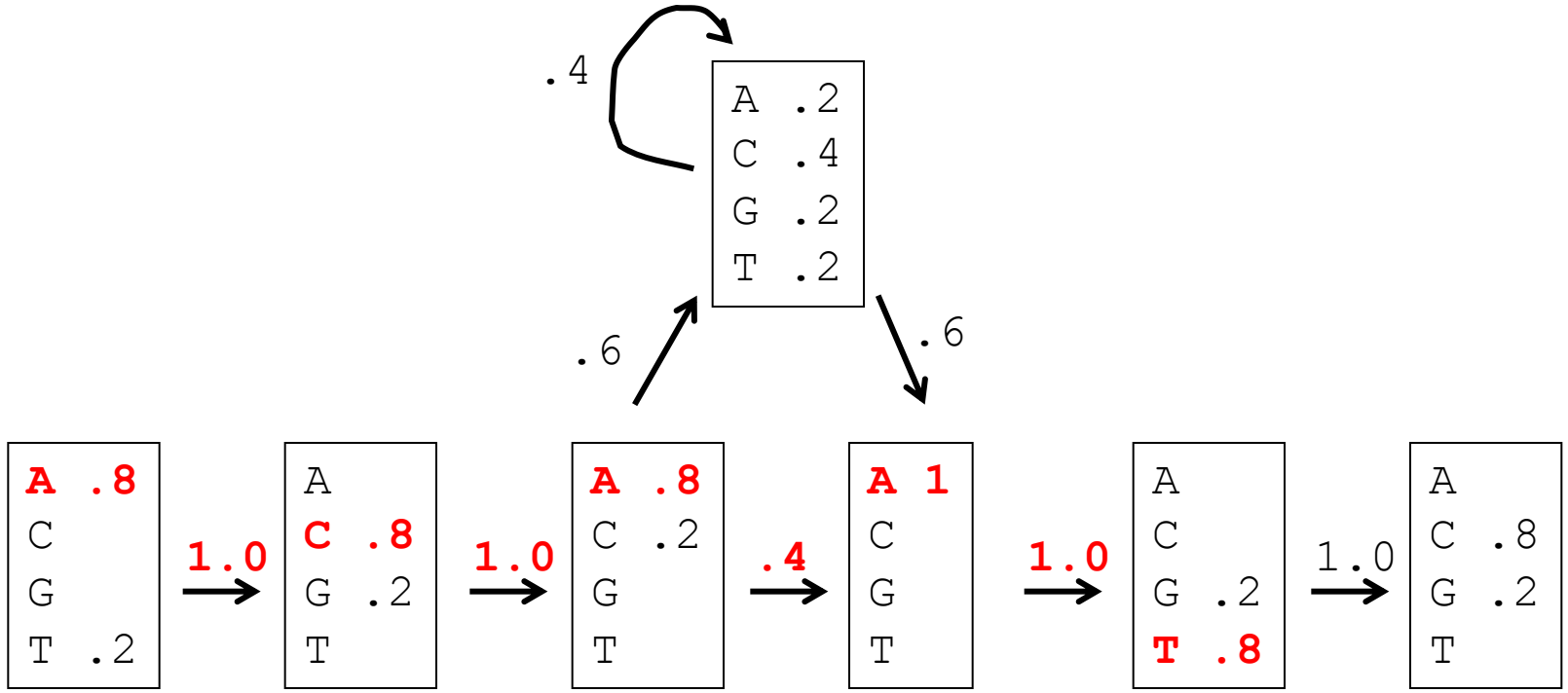
```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

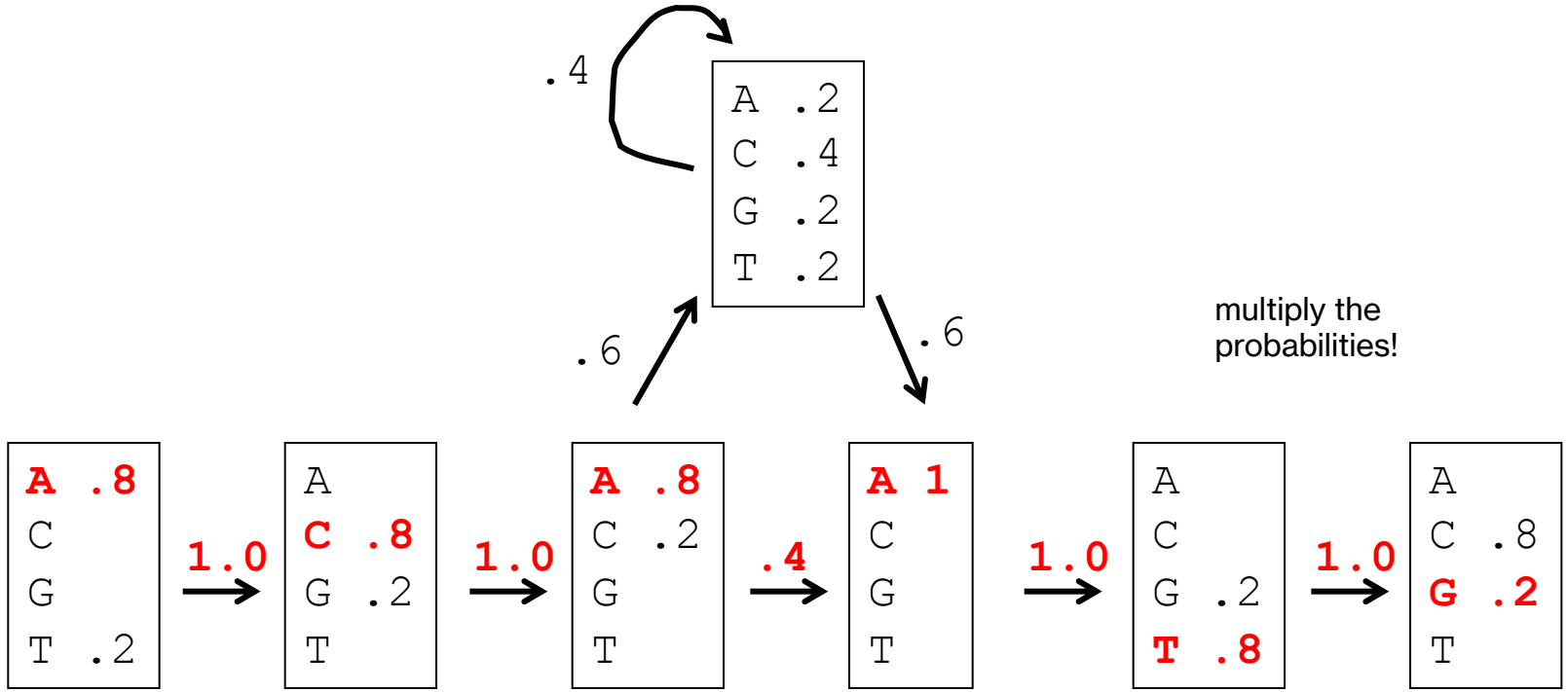


```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```



```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

P(ACAATG)=0.0328

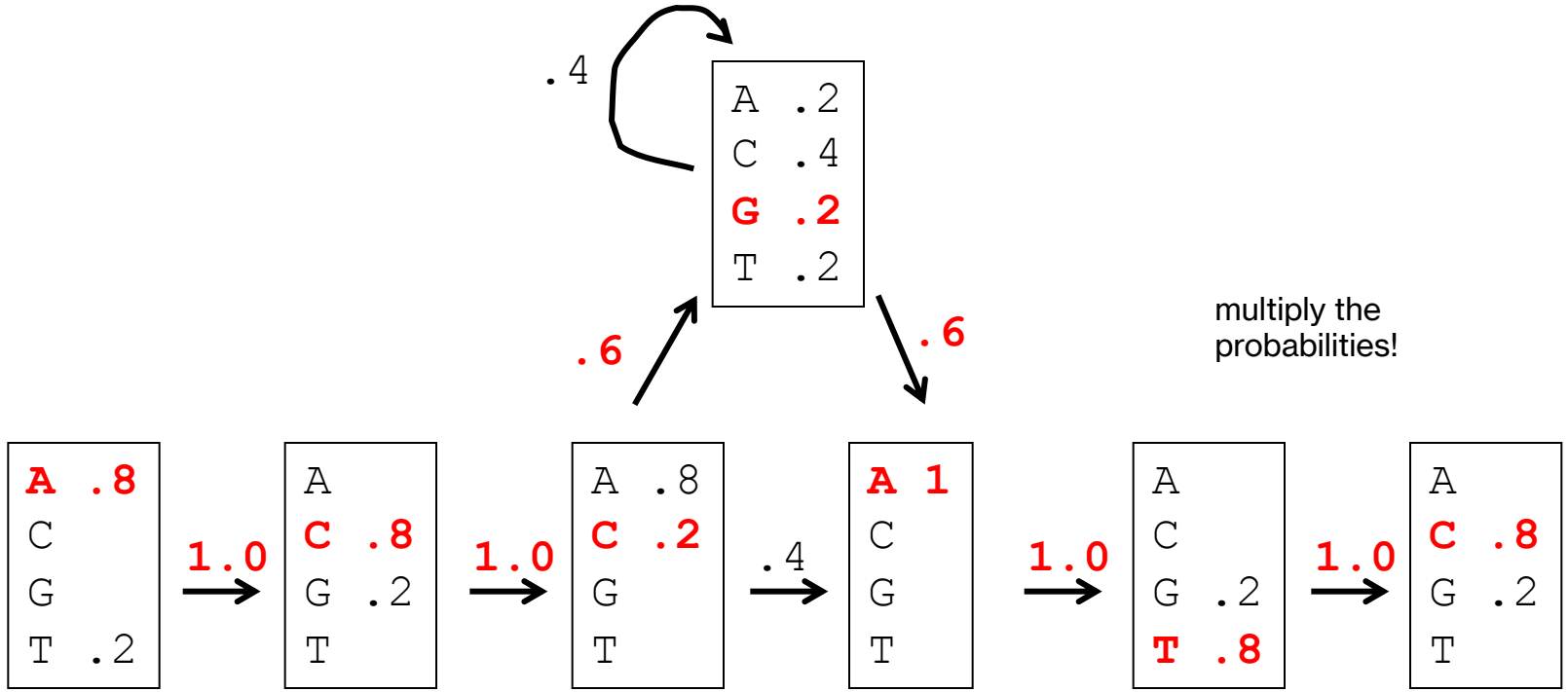


multiply the probabilities!

```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

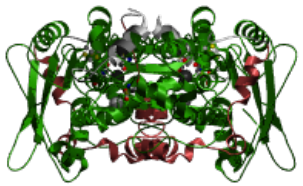
P(ACAATG)=0.0328

P(ACCGATC)=0.006



multiply the probabilities!

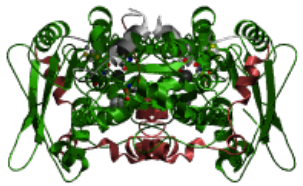
Pfam Iron Hydrogenase HMM



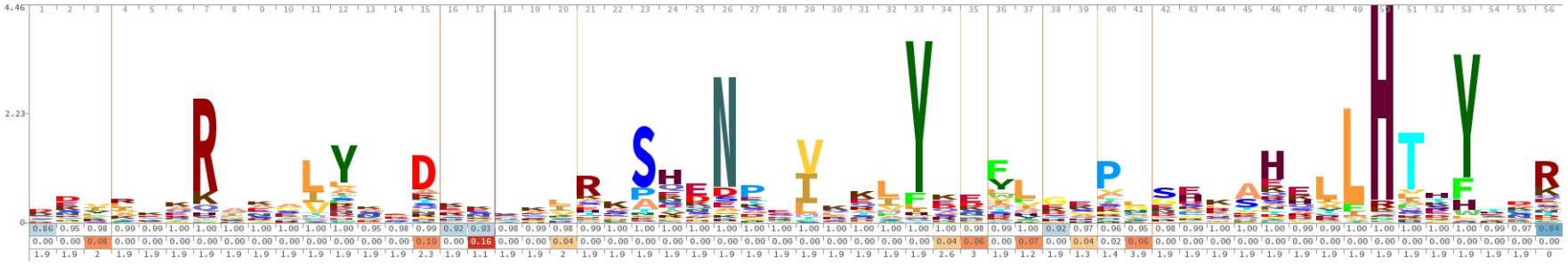
- 334 iron hydrogenases in the seed alignment

```
D9QTQ6_ACEAZ/525-580      MDV--KAKRAEALYQTD---KA-NTI--RKSHENPQIIKLYE---D----YL-GE-P-L-----SSDSHHLLHTSYQER
E3DPJ1_HALPG/515-570      NEK--KEKRGSGLSNID---DS-SKI--RKSHENPQIIKLYE---E----FL-GE-P-L-----GGESHHLLHTKYKAR
E4RJ60_HALHG/516-571      YEK--KVKRGVGLSGID---DK-SAV--RKSHKNPQVIKLYK---E----FL-GK-P-L-----SGESHHLLHTTYKSR
Q0AVN1_SYNWW/44-102       DDY--IAKRAAGLYTLD---ES-MAI--RKSHENPEVIQIYQ---D----FL--S-P-GKLECVSPKAHHLLHTKYGQ-
D7CNL1_SYNLT/44-102       DDY--IAKRAQGLYTLD---EK-MTI--RKSHENPEIIQLYK---D----FL--S-P-GEVKPMSEKAHHLLHTRYGQ-
L0KCP1_HALHC/517-572      KEI--KAKRGQGLYNID---QS-DKI--RKSHENPEIKKLYE---D----FL-GA-P-L-----SEKAHHLLHTNYQKR
R5AAQ0_9FIRM/470-524      EEL--YGVRGERLYTLD---AE-NPM--RFAHENPEVQALYH---E----YL-GE-P-L-----GETAHHLLHTDHKA-
F7V3Y0_CLOSS/515-570      KEM--AASRAPILYAFD---QI-TDL--RFSHENPSITKVYS---E----YL-GE-P-L-----SEKSHHLLHTDHHAW
R7B2L8_9BACE/516-570      QEL--AKDRAPILYSLD---RS-KNI--RFSHENPDVLKMYE---E----FF-EK-P-N-----SPVAHKLLHTDHHA-
R5TQ56_9FIRM/517-570      -EL--ADVRGRNLYKLD---KK-NPL--RFSHENPSVIKAYE---D----FF-EK-P-L-----SHKSHELLHTDHEA-
R6G054_9FIRM/519-572      KEM--AEIRSKNLYFLD---SQ-NER--RFSHENPEVLKTYE---E----YL-EK-P-L-----SRMSHKLLHTDHH--
F4GHP6_SPHCD/517-571      -EL--ASTRADVLYGLD---KV-DNL--RFSHENPSVLKAYE---S----FF-GK-P-L-----GHKCHELLHTDHHAW
R6K3U5_9FIRM/237-290      -DK--VAERCKVLYGLD---KV-NNV--RFSHENPEVLQCYR---D----YF-KE-P-L-----SEKSHELLHTSHTV-
R7BDK2_9FIRM/518-573      VEM--AADRAKELYKLD---KN-KQI--RFSHCNPEIHTIYK---E----YF-GK-P-L-----SPVSHHLLHTDHKYR
R6A2A7_9ACTN/541-595      -EL--AAERGQVLWGLD---AK-ADI--RFSHENPGVQACYR---E----FL-GA-P-L-----SPLAEELLHTDHHAW
R7D1R5_9ACTN/521-575      VEL--ADERAAVLRALD---HD-AQI--RFSHENPDVAACYR---D----FL-GE-P-L-----SELSEKLLHTDHTA-
G4KSU3_OSCVS/517-571      VEM--AAERGELLWELD---AK-SKI--RFSHENPDIKTLYS---E----YL-KE-P-L-----GKKSHHLLHTDHAA-
R6GQ90_9FIRM/517-572      VEL--AEKRGSVLWSID---KA-SPC--RFSHENPDVRELYR---D----YL-KK-P-L-----SDVSHHLLHTDHQAW
R5D0I3_9FIRM/517-570      TEM--AEARGNVLWSID---KK-SPV--RFSHENPEVQTLYR---E----YL-RA-P-L-----SGRSHHLLHTDHE--
R6Q6Y7_9FIRM/517-570      --Q--AERRGNHLYFLD---DI-ANL--RFSHENPAIQALYK---N----FL-GE-P-L-----GEKAHHLLHTDHTAW
R6RTM0_9FIRM/518-571      QEL--AEERGSSLYFLD---RD-TEI--RFSHDNPDIQNLYE---E----FF-EK-P-L-----SHRAHQLLHTEHQ--
Q73MB6_TREDE/520-573      GEL--AVKRGSNLYFID---KN-SKV--RYSHENECIKALYN---D----FF-EK-P-N-----SHKAHSLLHTDHF--
R5J469_9CLOT/520-573      -EM--AFERGKNLYFLD---EN-ADI--RRSHENPDVKALYD---N----YF-EQ-P-L-----SHKSHMLLHTDHNK-
D4M4H0_9FIRM/519-573      EEL--AHTRGANLYFLD---KN-AKI--RFSHENQDVMKLYN---D----FL-EK-P-L-----SHKSHMLLHTDHTK-
R6LI15_9FIRM/519-573      EEL--ARTRGENLYFLD---KN-APL--RFSHENPDVLRLYR---D----FF-EK-P-L-----SHKSHMLLHTDHNA-
F0T2S8_SYNGF/456-511      DTI--RTQRSNSLYTLD---KN-AKV--RNSHENTEITQIYK---D----YL-HA-P-M-----SHLAEEILHTEYESR
A5D4I9_PELTS/463-518      DTV--REQRLAALYKAD---ASLSK---RKSYENEEVAALYR---D----FL-GH-P-M-----SELAEELLHTEYHSR
R4KFN4_9FIRM/459-514      DEV--RMQRINSLYQAD---AR-AQR--RESHENAEVLALYQ---N----FL-KH-P-M-----SELAEELLHTKYTDR
F6DPY8_DESRL/455-511      DQV--RQARLNSLYTMD---AKMYKK--RLSHENSEVLQLYK---N----YL-EQ-P-M-----SHLAEELLHTEYTDR
```

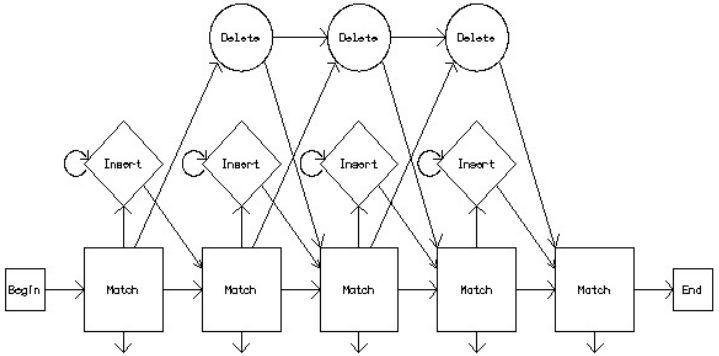
Pfam Iron Hydrogenase HMM



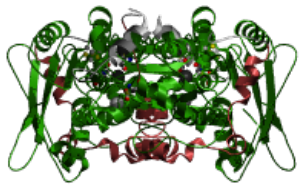
- Much sequence variation, but some conservation



- HMMER / Pfam has a generalized HMM that can be trained by any protein domain



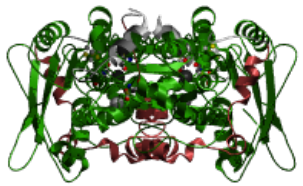
Pfam Iron Hydrogenase HMM



- The HMM is trained using the seed alignment to determine the emission and transition probabilities

HMM	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R
	m->m	m->i	m->d	i->m	i->i	d->m	d->d								
COMPO	2.66299	4.87199	2.91679	2.55591	3.66715	3.29500	2.87691	3.09550	2.45416	2.39490	3.84919	3.01061	3.43251	3.03167	2.71
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89
	0.15412	5.62693	1.97159	0.61958	0.77255	0.00000	*								
1	2.69450	5.20126	1.90404	2.47178	4.68913	3.78335	4.01145	3.75163	2.20418	2.96186	3.33300	2.43140	3.49632	2.59266	3.14
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89
	0.00718	5.47902	5.81517	0.61958	0.77255	0.42007	1.07002								
2	2.91083	4.82906	1.58312	1.73987	5.01707	3.82147	3.64171	4.50313	2.01029	3.15122	4.59410	2.83954	4.25294	3.15745	3.12
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89
	0.00559	5.58603	6.30838	0.61958	0.77255	0.42789	1.05522								
3	2.63456	4.70051	4.84319	2.98389	2.64502	4.37715	4.37000	1.96988	3.23179	2.30513	3.40557	3.77117	4.64207	4.19062	3.25
	2.68582	4.42236	2.77514	2.73093	3.46365	2.40524	3.72482	3.29365	2.67744	2.69351	4.24701	2.90358	2.73751	3.18157	2.89
	0.08136	2.57205	6.34224	0.67760	0.70894	0.47041	0.98016								
4	2.54801	4.72728	4.74569	4.15699	3.82991	4.36365	4.46123	1.93633	2.71591	1.80176	2.81197	4.19038	4.73278	4.00802	1.74
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89
	0.00537	5.62610	6.34845	0.61958	0.77255	0.48306	0.95944								
5	2.11055	5.19463	3.13477	2.25423	4.84332	3.10661	3.88848	3.85885	1.98083	2.26284	3.93265	3.22775	4.35588	2.50230	2.73
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89
	0.00536	5.62693	6.34927	0.61958	0.77255	0.48576	0.95510								
6	2.03322	4.91437	2.94559	2.42989	5.02036	3.66309	4.04897	3.63439	1.52459	2.77931	3.89988	3.20092	4.15513	2.30684	2.83
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89
	0.00536	5.62693	6.34927	0.61958	0.77255	0.48576	0.95510								
7	4.73891	6.64548	5.68057	4.42078	6.41719	5.15128	3.93502	5.55175	2.29408	4.78890	5.71680	4.63644	5.41712	2.87234	0.31
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89
	0.00536	5.62693	6.34927	0.61958	0.77255	0.48576	0.95510								
8	1.62426	4.44842	4.26891	3.69735	3.86164	3.13257	4.52444	2.32789	3.39179	2.57955	2.66032	3.91731	4.63345	2.24662	2.85
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89
	0.00536	5.62693	6.34927	0.61958	0.77255	0.48576	0.95510								

Pfam Iron Hydrogenase HMM



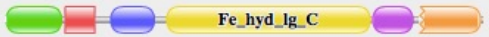
- The HMM is trained using the seed alignment to determine the emission and transition probabilities

- Pfam is a collection of many trained HMMs; query sequences are compared to all of Pfam to find the best fitting HMMs

Sequence search results

Show the detailed description of this results page.

We found **6** Pfam-A matches to your search sequence (**all** significant)



Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

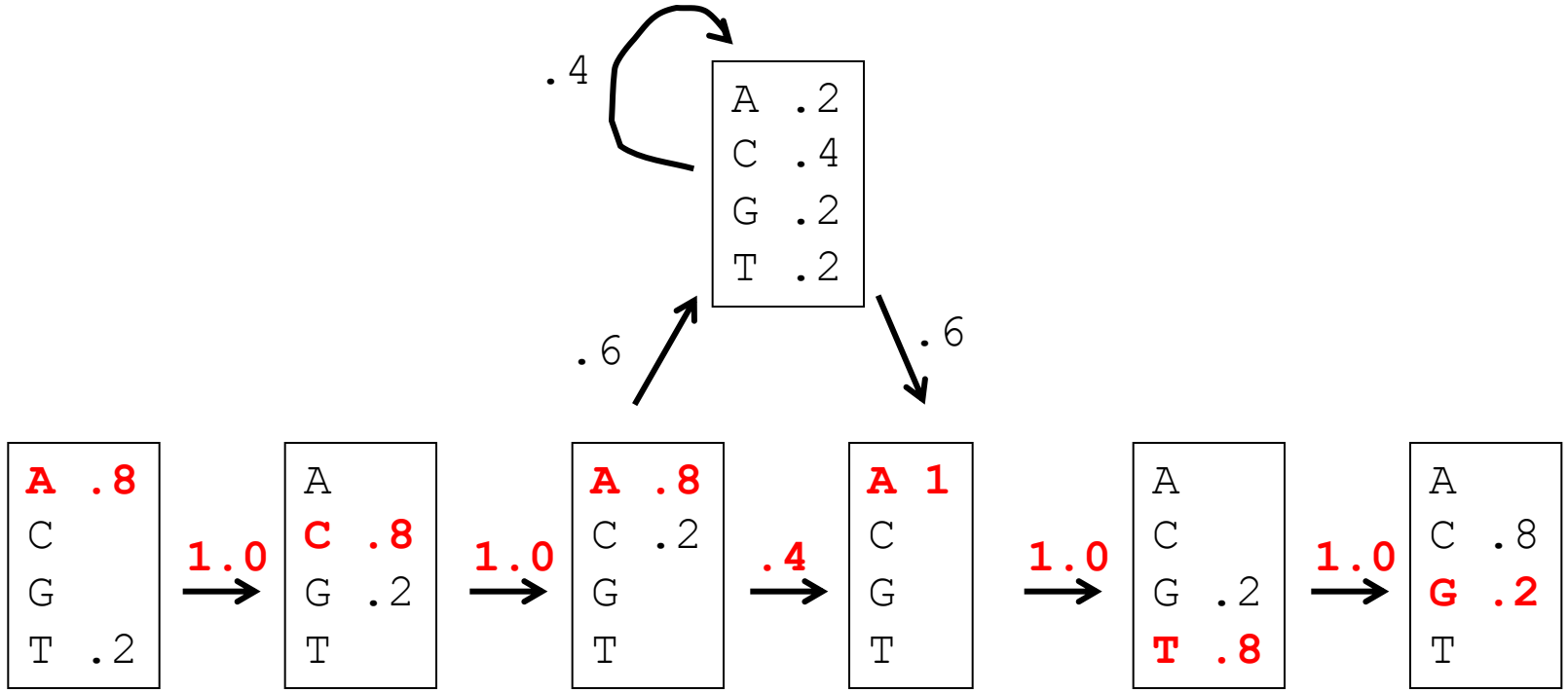
Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value
				Start	End	Start	End	From	To			
Fer2_4	2Fe-2S iron-sulfur cluster binding domai ...	Domain	CL0486	1	75	2	74	**5**	**81**	82	45.5	5.3e-12
NADH-G_4Fe-4S_3	NADH-ubiquinone oxidoreductase-G iron-su ...	Domain	n/a	81	120	81	120	1	40	40	68.0	3.4e-19
Fer4_7	4Fe-4S dicluster domain	Domain	CL0344	142	200	142	200	1	52	52	31.5	1.8e-07
Fe_hyd_lg_C	Iron only hydrogenase large subunit, C-t ...	Domain	n/a	218	493	218	493	1	248	248	286.8	1.3e-85
Fe_hyd_SSU	Iron hydrogenase small subunit	Domain	n/a	498	551	500	551	**3**	56	56	29.5	5.1e-07
2Fe-2S_thioredx	Thioredoxin-like [2Fe-2S] ferredoxin	Family	CL0172	563	644	569	641	**65**	**141**	145	34.8	1.2e-08

How does searching work?

What does Markov mean?

What is Hidden?

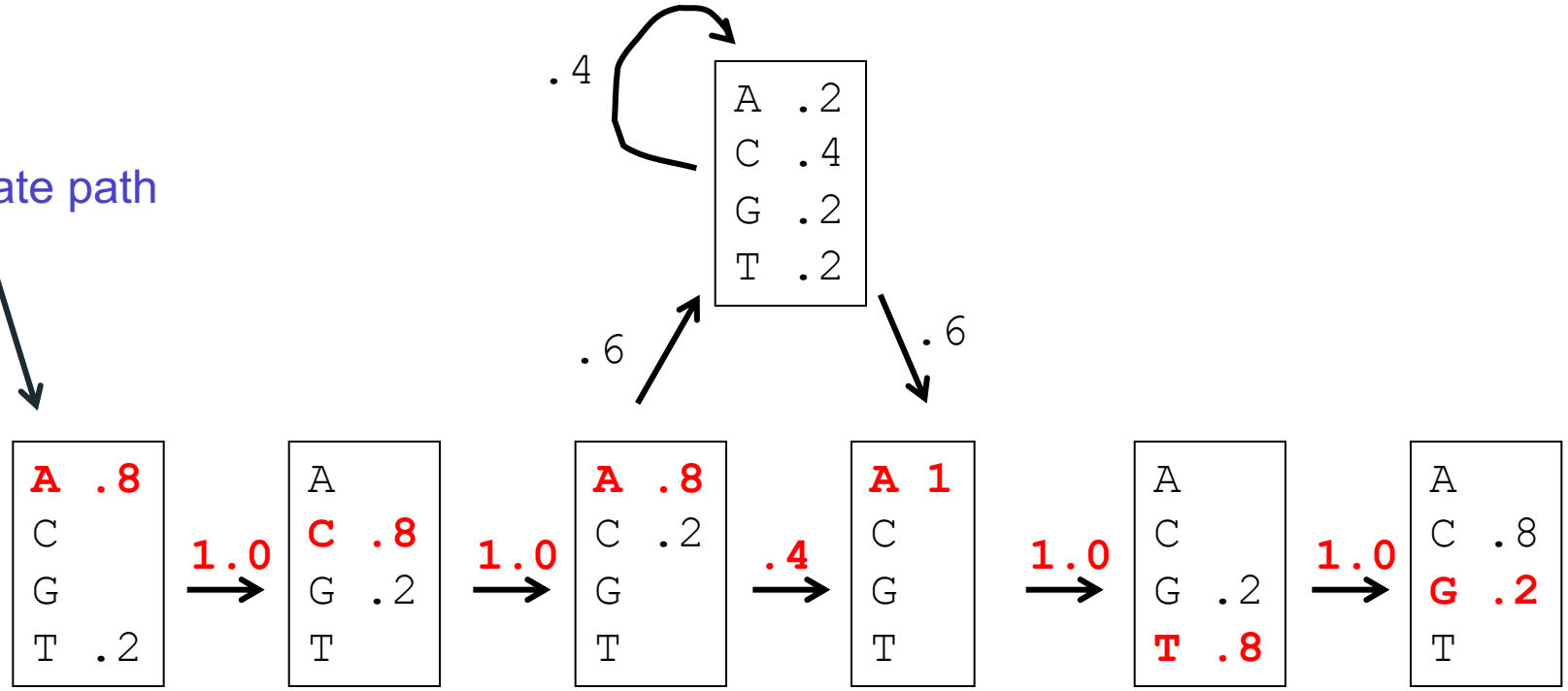
ACA---ATG



observed sequence

ACA---ATG

state path



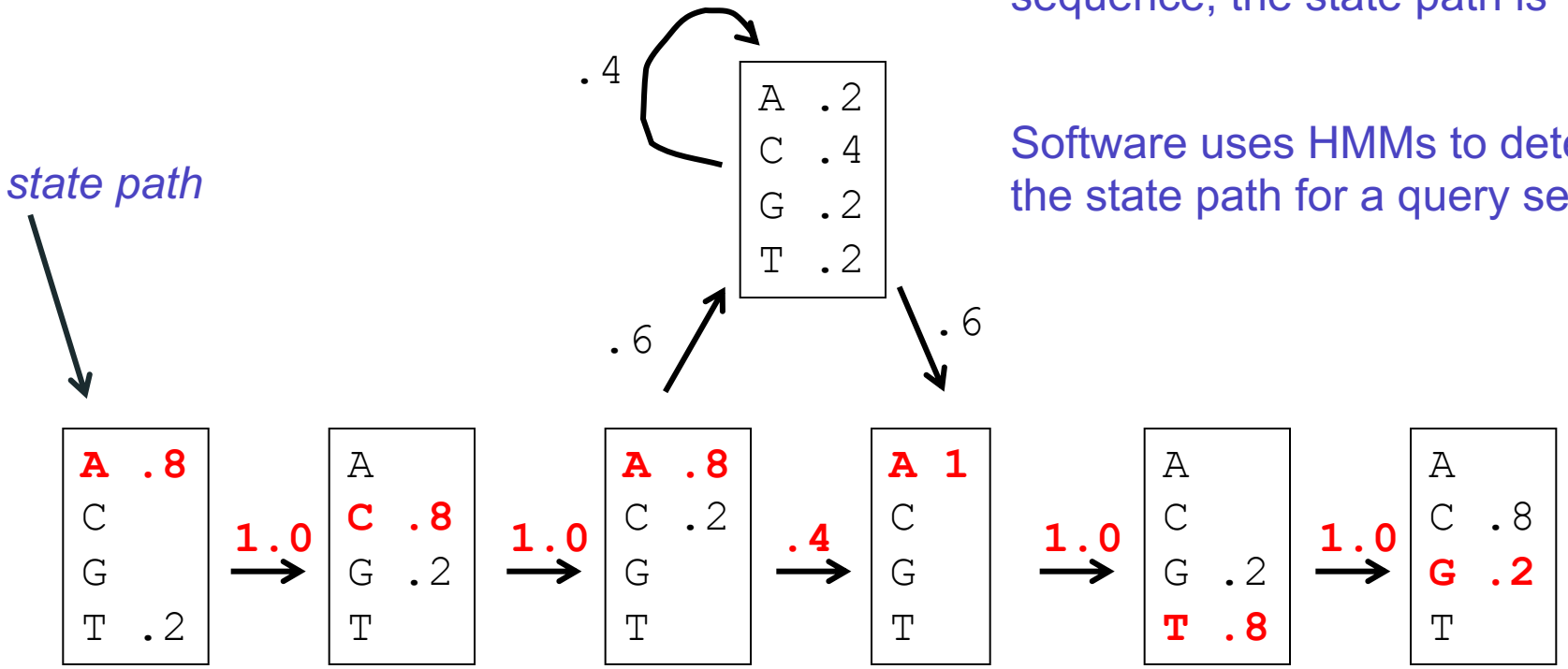
observed sequence

ACA---ATG

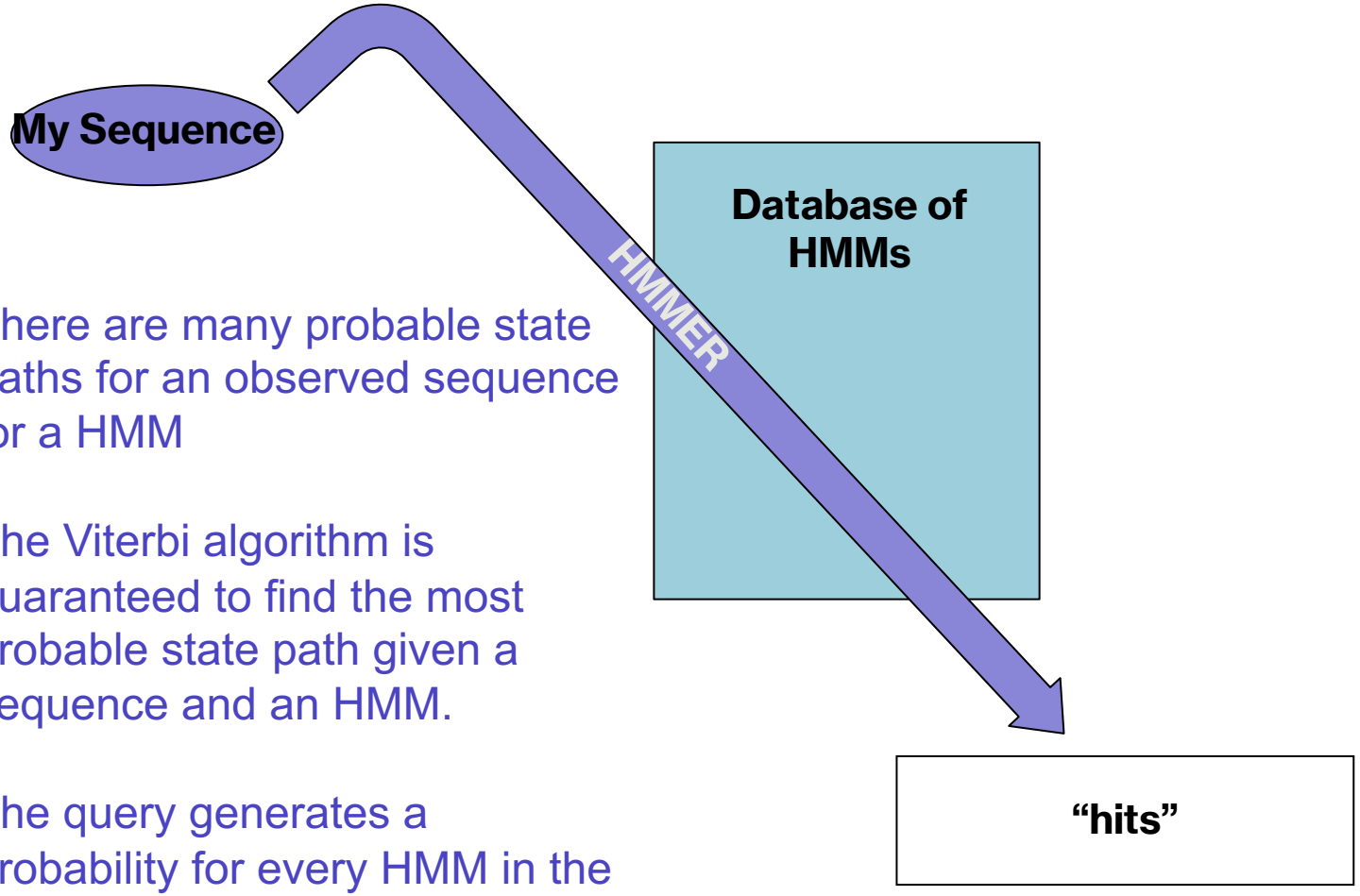
Markov Chain – next state emission depends only upon current state, i.e. the model is a linear chain

Since we only know our query sequence, the state path is “hidden”

Software uses HMMs to determine the state path for a query sequence



Hidden Markov Models (HMMs)



There are many probable state paths for an observed sequence for a HMM

The Viterbi algorithm is guaranteed to find the most probable state path given a sequence and an HMM.

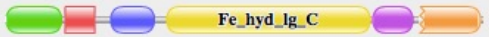
The query generates a probability for every HMM in the database

HMMs in Pfam / Hmmer

Sequence search results

Show the detailed description of this results page.

We found **6** Pfam-A matches to your search sequence (**all** significant)



Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value
				Start	End	Start	End	From	To			
Fer2_4	2Fe-2S iron-sulfur cluster binding domai ...	Domain	CL0486	1	75	2	74	**5**	**81**	82	45.5	5.3e-12
NADH-G_4Fe-4S_3	NADH-ubiquinone oxidoreductase-G iron-su ...	Domain	n/a	81	120	81	120	1	40	40	68.0	3.4e-19
Fer4_7	4Fe-4S dicluster domain	Domain	CL0344	142	200	142	200	1	52	52	31.5	1.8e-07
Fe_hyd_lg_C	Iron only hydrogenase large subunit, C-t ...	Domain	n/a	218	493	218	493	1	248	248	286.8	1.3e-85
Fe_hyd_SSU	Iron hydrogenase small subunit	Domain	n/a	498	551	500	551	**3**	56	56	29.5	5.1e-07
2Fe-2S_thioredx	Thioredoxin-like [2Fe-2S] ferredoxin	Family	CL0172	563	644	569	641	**65**	**141**	145	34.8	1.2e-08

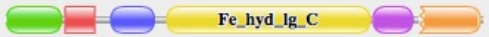
- The HMMER software uses the probability values to calculate:
  - bit score – a log-odds score of the fit of the query to the HMM; the higher the score, the better the alignment
  - expectation value - estimates the likelihood that a given match is purely by chance; a function of database size
- Pfam website uses a default expectation value cut-off of 1.0

HMMs in Pfam / Hmmer

Sequence search results

Show the detailed description of this results page.

We found **6** Pfam-A matches to your search sequence (**all** significant)



Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value
				Start	End	Start	End	From	To			
Fer2_4	2Fe-2S iron-sulfur cluster binding domai ...	Domain	CL0486	1	75	2	74	**5**	**81**	82	45.5	5.3e-12
NADH-G_4Fe-4S_3	NADH-ubiquinone oxidoreductase-G iron-su ...	Domain	n/a	81	120	81	120	1	40	40	68.0	3.4e-19
Fer4_7	4Fe-4S dicluster domain	Domain	CL0344	142	200	142	200	1	52	52	31.5	1.8e-07
Fe_hyd_lg_C	Iron only hydrogenase large subunit, C-t ...	Domain	n/a	218	493	218	493	1	248	248	286.8	1.3e-85
Fe_hyd_SSU	Iron hydrogenase small subunit	Domain	n/a	498	551	500	551	**3**	56	56	29.5	5.1e-07
2Fe-2S_thioredx	Thioredoxin-like [2Fe-2S] ferredoxin	Family	CL0172	563	644	569	641	**65**	**141**	145	34.8	1.2e-08

- Pfam HMMs are curated by functional biologists and model functional domains; Pfam HMMs are thus one of the most powerful tools available for prediction of protein function

- Caveat – correlations among residues cannot be modeled by HMMs as Markov Chains cannot ‘remember’ earlier states; secondary structure not workable in HMMs

What about finding short sequences?

- Short sequences or motifs by definition have less information
  - Often too short for BLAST (below word size or extension rules)
  - Not enough information to build an HMM
  - Easily match random sequences – expectation values break down
  - Statistical confidence and avoidance of false discovery difficult for reasons listed above – experimental validation often required

- Two common bioinformatics questions
  - Detection of amino acid motifs, e.g. PROSITE database
  - Detection of DNA binding sites, e.g. JASPAR database

- Two common methods
  - Pattern matching
  - Position-specific scoring matrix (PSSM)

What about finding short sequences?

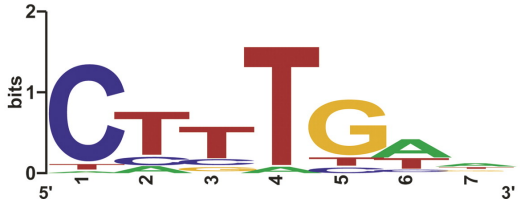
- Pattern matching, e.g. C-x-H-x-[LIVMFY]-C-xx-C-[LIVMYA]
  - Not statistical – the pattern exists in the subject or not
  - Frequently important for analysis of proteins (e.g. PROSITE)
  - Computers are very good at pattern matching – fast!
  - Universal language for pattern matching – Regular Expressions (RegEx)
  - Almost exclusively a command-line tool

- Position-specific scoring matrix (PSSM)
  - A pattern with variation based on observation
  - Also important for analysis of proteins (e.g. PROSITE)
  - Particularly important for analysis of DNA binding sites (e.g. JASPAR database)
  - Commonly generated from ChIP-Seq results
  - Meme/Mast software suite and other suites at the command line
  - Statistical in nature, but very high false discovery rate!

A

	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9

B



This Week...

WEEK 3 (SEPTEMBER 20 and 22) - SEQUENCE SIMILARITY & SEARCHING

LIVE Class update on Wednesday,

Recorded Content

- Lecture #2 - Sequencing Similarity & Searching,
- Dr. Joanna Wilson - The Shark CYPome, https://web.microsoftstream.com/video/a876db13-6d45-4ac0-86c5-5c0ef83496e6
- Overview & Demo of Laboratory #2 - Protein Annotation & Gene Finding, https://web.microsoftstream.com/video/b0eb4084-0452-479e-a33b-556abd9809bc

Tutorial

- **LIVE** session with Teaching Assistants
- Tutorial content can be found at GitHub, answers due on A2L
    - Monday,
    - Wednesday,

Flash Updates

- **BLAST**. Provide a review of the purpose of BLAST algorithms for database searching and how to perform them online. Specifically, outline the difference between BLASTN, BLASTP, BLASTX, TBLASTN, and TBLASTX. See Lobo 2008. Basic Local Alignment Search Tool (BLAST). Nature Education 1: 215 [http://www.nature.com/scitable/topicpage/basic-local-alignment-search-tool-blast-29096]
- **Pfam**. Provide a review of the Pfam resource, with an emphasis on the variety of tools and data it offers. See Nucleic Acids Res. 2019 Jan 8;47(D1):D427-D432 [PMID 30357350] and Nucleic Acids Res. 2018 Jul 2;46(W1):W200-W204 [PMID 29905871].
- **PROSITE**. Provide a review of the PROSITE resource, with an emphasis on the variety of tools and data it offers. See Nucleic Acids Res. 2013 41(Database issue):D344-7 [PMID 23161676].