Tools and Algorithms in Bioinformatics

GCBA815/MCGB815/BMI815, Fall 2017

Week-4: Profiles, HMMs, Advanced BLAST

Babu Guda, Ph.D.
Professor, Genetics, Cell Biology & Anatomy
Director, Bioinformatics and Systems Biology Core
University of Nebraska Medical Center

Different types of BLAST programs

Program	Query	Database	Comparison
blastn	nucleotide	nucleotide	nucleotide
blastp	protein	protein	protein
blastx	nucleotide	protein	protein
tblastn	protein	nucleotide	protein
tblastx	nucleotide	nucleotide	protein
megablast	nucleotide	nucleotide	nucleotide
PHI-Blast	protein	protein	protein
PSI-Blast	Protein-Profile	protein	protein
RPS-Blast	protein	Profiles	protein

Different types of alignments

Local Alignment

Pairwise Sequence Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

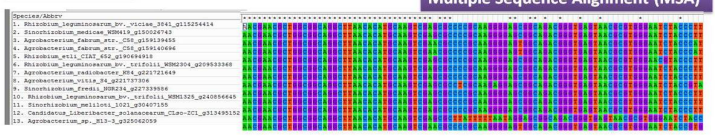
Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||||    |||||||  ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

Multiple Sequence Alignment (MSA)



Source: www.majordifferences.com

Multiple alignments of Voltage-Gated Ion Channel (VGC) Protein Families

Sodium VGC

```
YNMTAEFEEMLQVGNLVFTGIFTAEMTFKIIALD
HPMTPQFEHVLAVGNLVFTGIFTAEMFLKLIAMD
FPMSEELRSLLHVGNLVFIGIYTIELILKIIAMH
YPMTEHFDNVLSVGNLVFTGIFTAEMVLKLIAMD
YPMTEEFNNVLNVGNLVFTGIFTAEMVLKLIALD
YPMTEEFDYMLSVGNLVFTGIFAAEMFFKLIAMD
DEQSAEKTKILNKINQFFVAVFTGECVMKMFALR
DDQSEEKTKILGKINQFFVAVFTGECVMKMFALR
DNQSEEKTKVLGRINQFFVAVFTGECVMKMFALR
DEQGEEKTKVLGRINQFFVAVFTGECVMKMFALR
YNQPKAMKSILDHLNWVFVVIFTLECLIKIFALR
EGQPNEVKKIFDILNIVFVVIFTVECLIKVFALR
ADQPKDVKKTFDILNIAFVVIFTIECLIKVFALR
DDQSEYVTTILSRINLVFIVLFTGECVLKLISLR
```

Calcium VGC

```
HYYFTNSWNIFDFVVVILSIVGTVLSDIIQ-
HYYFTVGWNIFDFVVVIISIVGMFLADLIE-
HYYFTIGWNIFDFVVVILSIVGMFLAELIE-
HYYFTVGWNIFDFVVVILSIVGMFLAEMIE-
HYYFTIGWNIFDFVVVILSIVGMFLAELIE-
LEYFKYGMNVFDFVIVVFSIAVIIMIEYDE-
WFYFKDPWNVFDFSVVVFSTIAWILQFFES-
FYYFKNPWNVFDFIVVILSVVGSTMNEVIK-
WYYFKEPWNVFDCVVVTLSILGIAIKDLIA-
```

Potassium VGC

```
LRVIRLVRVFRIFKLSRHSKGLQI
LRVIRLVRVFRIFKLSRHSKGLQI
LRVIRLVRVFRIFKLSRHSKGLQI
VQVFRIMRILRVLKLARHSTGLQS
LKVVRLLRLGRVVRKLDRYLEYGA
FVTLRVFRVFRIFKFSRHSQGLRI
LEFFSIIRIMRLFKLTRHSSGLKI
VQIFRIMRILRILKLARHSTGLQS
VQIFRIMRILRILKLARHSTGLQS
VQIFRIMRILRILKLARHSTGLQS
VQIFRIMRILRILKLARHSTGLQS
```

Multiple Sequence Alignments

- Aligning homologous residues among a set of sequences, together in columns
- Multiple alignments exhibit structural and evolutionary information among closely related species i.e., families
    - A column of aligned residues in a conserved region is likely to occupy similar 3-D positions in protein structure
- Patterns or motifs common to a set of sequences may only be apparent from multiple alignments
- Necessary for building family profiles, phylogenetic trees and extracting evolutionary relationships
- Useful for homology modeling if at least one member of the family has a known 3-D structure

Scoring a Multiple Alignment

- Some positions are more conserved than others
- Sequences are not independent, but are related by a phylogeny

```
   430667           100.0%   MTDREQFFRDVNKIVIKIGTSSITRKGCDHTKENCNIDPAFMESIAAQVY
 1 SWALL:PROB_MEIRU  38.4%   ------------LPLTAQTHRRLVVKVGSAVLSGPQGRQHQLAIAAQVA
 2 SWALL:Q9KPT8      38.4%   MTTNQRSIKAPQTVVVKLGTSVLTGG-------TLALDRAHMVELARQCA
 3 SWALL:AAG07953    36.5%   ---MRDKVTGARRVVVKIGSALLTADGR-------GLDRNAMAVVVEQMV
 4 SWALL:Q9RTD8      38.6%   -----------MRVVLKLGTSVLTA-GTDRLHRPRLVD--LMRDIAA---
 5 SWALL:PROB_THETH  34.9%   -------------RVEEGVLIPRPETEGLVELALGLPLPPAPRIARQVA
 6 SWALL:PROB_HAEIN  38.0%   --------MNKKTIVVKFGTSTLTQGS---PKLN---SPHMME-IVRQIA
 7 SWALL:PROB_ECOLI  36.6%   -------MSDSQTLVVKLGTSVLTGG-------SRRLNRAHIVELVRQCA
 8 SWALL:Q9KCR4      37.1%   --------MKRQRIVIKIGSSSLT------TKQGA-LDLEKLEGYVRAIV
 9 SWALL:PROB_SYNY3  36.1%   ----MTMAMQPQTLVIKIGTSSLARP------ETGQLALSTIAALVETVC
10 SWALL:PROB_BACSU  37.2%   --------MKKQRIVVKIGSSSLTnkG--------SIDEAKIREHVQAIS
11 SWALL:PROB_SERMA  35.3%   -------MNGSQTLVVKLGTSVLTGG-------SLRLNRAHIVELVRQCA
12 SWALL:O07509      34.3%   ----MTPDTSMKRVVVKIGSSSLTSL-------HGEISIRKLEALVDQVV
13 SWALL:Q9RDJ9      34.7%   MAGARQAVGEARRIVVKVGSSSLTTAAG-------GLDADRVDALVDVLA
14 SWALL:PROB_MYCTU  33.5%   RSPHRDAIRTARGLVVKVGTTALTTP-------SGMFDAGRLAGLAEAVE
15 SWALL:PROB_AQUAE  35.0%   -----------MRIVFKIGSNLL------ETDEG-DIDLSFLSKLAEGIK
```

Pair-wise

```
LVD--LMRDIAA----VSAQGHE
GLPLPPAPRIARQVAALREEGRE
```

Multiple

```
KIGTSSITRKGCDHTKENCNIDPAFME
TAQTHRRLVVKVGSAVLSGPQGRQHQL
KLGTSVLTGG-------TLALDRAHMV
KIGSALLTADGR-------GLDRNAMA
KLGTSVLTA-GTDRLHRPRLVD--LMR
```

• Most of the multiple alignment algorithms assume that the individual columns of an alignment are statistically independent

$$\text{Total Alignment Score } S = \Sigma_i S(m_i) + G$$

where, m_i is column i of the multiple alignment m,

$S(m_i)$ is the score for column i,

G is gap penalty

Table 1 - The log odds matrix for 250 PAMs (multiplied by 10)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2	-2	0	0	-4	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3
C		12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0
D			4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4
E				4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-7	-4
F					9	-5	-2	1	-5	2	0	-4	-5	-5	-4	-3	-3	-1	0	7
G						5	-2	-3	-2	-4	-3	0	-1	-1	-3	1	0	-1	-7	-5
H							6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0
I								5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1
K									5	-3	0	1	-1	1	3	0	0	-2	-3	-4
L										6	4	-3	-3	-2	-3	-3	-2	2	-2	-1
M											6	-2	-2	-1	0	-2	-1	2	-4	-2
N												2	-1	1	0	1	0	-2	-4	-2
P													6	0	0	1	0	-1	-6	-5
Q														4	1	-1	-1	-2	-5	-4
R															6	0	-1	-2	2	-4
S																2	1	-1	-2	-3
T																	3	0	-5	-3
V																		4	-6	-2
W																			17	0
Y																				10

Profile Analysis

- A profile contains position-specific weights for amino acid substitution and position specific penalties for gaps (insertions and deletions)
- A profile is a mathematical model for a group of sequences (families)

```
   430667           100.0%   MTDREQFFRDVNKIVIKIGTSSITRKGCDHTKENCNIDPAFMESIAAQVY
 1 SwALL:PROB_MEIRU  38.4%   ------------LPLTAQTHRRLVVKVGSAVLSGPQGRQHQLAIAAQVA
 2 SwALL:Q9KPT8      38.4%   MTTNQRSIKAPQTVVVKLGTSVLTGG-------TLALDRAHMVELARQCA
 3 SwALL:AAG07953    36.5%   ---MRDKVTGARRVVVKIGSALLTADGR-------GLDRNAMAVVVEQMV
 4 SwALL:Q9RTD8      38.6%   ----------MRVVLKLGTSVLTA-GTDRLHRPRLVD--LMRDIAA---
 5 SwALL:PROB_THETH  34.9%   -------------RVEEGVLIPRPETEGLVELALGLPLPPAPRIARQVA
 6 SwALL:PROB_HAEIN  38.0%   --------MNKKTIVVKFGTSTLTQGS---PKLN---SPHMME-IVRQIA
 7 SwALL:PROB_ECOLI  36.6%   -------MSDSQTLVVKLGTSVLTGG-------SRRLNRAHIVELVRQCA
 8 SwALL:Q9KCR4      37.1%   --------MKRQRIVIKIGSSSLT------TKQGA-LDLEKLEGYVRAIV
 9 SwALL:PROB_SYNY3  36.1%   ----MTMAMQPQTLVIKIGTSSLARP------ETGQLALSTIAALVETVC
10 SwALL:PROB_BACSU  37.2%   --------MKKQRIVVKIGSSSLTnkG--------SIDEAKIREHVQAIS
11 SwALL:PROB_SERMA  35.3%   -------MNGSQTLVVKLGTSVLTGG-------SLRLNRAHIVELVRQCA
12 SwALL:O07509      34.3%   ----MTPDTSMKRVVVKIGSSSLTSL-------HGEISIRKLEALVDQVV
13 SwALL:Q9RDJ9      34.7%   MAGARQAVGEARRIVVKVGSSSLTTAAG-------GLDADRVDALVDVLA
14 SwALL:PROB_MYCTU  33.5%   RSPHRDAIRTARGLVVKVGTTALTTP-------SGMFDAGRLAGLAEAVE
15 SwALL:PROB_AQUAE  35.0%   ----------MRIVFKIGSNLL------ETDEG-DIDLSFLSKLAEGIK
```


Multiple alignments from different VGC protein families

Sodium VGC

```
·YNMTAEFEEMLQVGNLVFTGIFTAEMTFKIIALD
·HPMTPQFEHVLAVGNLVFTGIFTAEMFLKLIAMD
·FPMSEELRSLLHVGNLVFIGIYTIELILKIIAMH
·YPMTEHFDNVLSVGNLVFTGIFTAEMVLKLIAMD
·YPMTEEFNNVLNVGNLVFTGIFTAEMVLKLIALD
·YPMTEEFDYMLSVGNLVFTGIFAAEMFFKLIAMD
·DEQSAEKTKILNKINQFFVAVFTGECVMKMFALR
·DDQSEEKTKILGKINQFFVAVFTGECVMKMFALR
·DNQSEEKTKVLGRINQFFVAVFTGECVMKMFALR
·DEQGEEKTKVLGRINQFFVAVFTGECVMKMFALR
·YNQPKAMKSILDHLNWVFVVIFTLECLIKIFALR
·EGQPNEVKKIFDILNIVFVVIFTVECLIKVFALR
·ADQPKDVKKTFDILNIAFVVIFTIECLIKVFALR
·DDQSEYVTTILSRINLVFIVLFTGECVLKLISLR
```

Calcium VGC

```
·HYYFTNSWNIFDFVVVILSIVGTVLSDIIQ-
·HYYFTVGWNIFDFVVVIISIVGMFLADLIE-
·HYYFTIGWNIFDFVVVILSIVGMFLAELIE-
·HYYFTVGWNIFDFVVVILSIVGMFLAEMIE-
·HYYFTIGWNIFDFVVVILSIVGMFLAELIE-
·LEYFKYGMNVFDFVIVVFSIAVIIMIEYDE-
·WFYFKDPWNVFDFSVVVFSTIAWILQFFES-
·FYYFKNPWNVFDFIVVILSVVGSTMNEVIK-
·WYYFKEPWNVFDCVVVTLSILGIAIKDLIA-
```

Potassium VGC

```
LRVIRLVRVFRIFKLSRHSKGLQI
LRVIRLVRVFRIFKLSRHSKGLQI
LRVIRLVRVFRIFKLSRHSKGLQI
VQVFRIMRILRVLKLARHSTGLQS
LKVVRLLRLGRVVRKLDRYLEYGA
FVTLRVFRVFRIFKFSRHSQGLRI
LEFFSIIRIMRLFKLTRHSSGLKI
VQIFRIMRILRILKLARHSTGLQS
VQIFRIMRILRILKLARHSTGLQS
VQIFRIMRILRILKLARHSTGLQS
VQIFRIMRILRILKLARHSTGLQS
```

Profile Analysis Procedure

• The starting point is a group of related sequences (probe) which are aligned by similarity in sequence or structure

• From the probe, a profile is made

• This profile is specific to the family of input sequences

• Once a profile is made, new sequences are progressively added to the existing ones to build multiple alignment

• The difference between a standard substitution score and a profile-based substitutions score is that in the latter, the scores are position-specific.

• Also, standard substitution matrices contain 'flat-rate' scores, while profile-based matrices are family- and position-specific.

PROFILEMAKE

• This program makes a profile by taking a family of aligned sequences as input

• A matrix M of 21 x N *(*total sequences*)* is needed to fill in the scores

• Score for amino acid j at position i is f_{ij} S_{ij}

• Total score for each column M_i is

$$M_i = \Sigma^{20}_{j=1} f_{ij} S_{jj}$$

where, f_{ij} is the frequency of residue j at position i in the aligned sequences (frequency is the ratio of no. of occurrences of residue j to N)

$S_{jj'}$ is the comparison score for residues j and j' in the basis scoring table (like PAM, BLOSUM etc)

• Position-specific penalties are calculated for insertions and deletions

POS	PROBE	CONSENSUS	PROFILE																				
			A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A	V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S	R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R	K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S	A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q	G	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S	F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L	T	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F F . L	F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D	S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S	S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N	A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T	Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y	D	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W	N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
.	.												.										
.	.												.										
.	.												.										
48	S G N S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y	S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

Advantages of Profile-based scores

- By incorporating position-specific information, profile-based scores can be customized to specific protein families to carry out highly sensitive and highly specific searches.
- Similarly, probes generated from structural information are extremely sensitive to identify structural patterns
- Profiles generated from sequences are specific for a protein family or superfamily. Significant similarity of a sequence to a profile suggests that the protein is homologous to that family or superfamily

PSI-BLAST

- Position-specific Iterated BLAST
- The output from regular BLAST search is converted into a multiple alignment using some heuristics
- A profile (probe) is developed using the above multiple alignment
- Using the above probe, a new BLAST search is done which picks up a new set of sequences
- The profile is updated by including the new sequences and a new probe is created
- This process is iterated until there are no more new hits (convergence)
- Usually, PSI-BLAST converges after 4-5 iterations

PHI-BLAST

- Pattern-Hit Initiated BLAST
- This program combines matching regular expressions with local alignments surrounding the match
- Given a protein sequence S and a regular expression pattern P occurring in S, PHI-BLAST searches for occurrence of P and also sequences homologous in the vicinity of P.
- PHI-BLAST preferable to other flavors of BLAST since it efficiently filters out the random hits from the real homologous sequences
- The syntax for patterns follows the rules of PROSITE

  Example

  - [PVLM] means one occurrence of P or V or L or M
  - X means any residue
  - X(3) means 3 positions in which any residue is allowed
  - X(2, 5) means any residue is allowed in 2 to 5 positions

RPS-BLAST

(Reverse Position-Specific Blast)

- In PSI-BLAST, a profile is iteratively built for the query sequence and used to search against each sequence in the database
- In RPS-BLAST, a sequence is compared against a position-specific matrix and hence the term reverse
- NCBI has developed this program to search against CCD (Conserved Common Domains), but RPS-BLAST can also be run as a stand-alone program against local databases
- http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

HMMs for biological sequences

- Hidden Markov models are statistical models that were initially developed for speech recognition.
- The most popular use of HMMs in molecular biology is as a 'probabilistic profile' of a protein family, which is called a profile HMM.
- Apart from this, HMMs are also used for multiple sequence alignment, gene prediction (ORF finding), and protein structure prediction
- Advantages are: statistically sound, no sequence ordering or gap penalties are required
- Limitations are: A large number of homologous sequences are required to get statistically reliable models

Stochastic modeling of biological sequences

For Example, Profile is a position-specific scoring matrix.

Family Members
Position 1 2 3 4 5
CCGTL
CGHSV
GCGSL
CGGTL
CCGSS

Position	1	2	3	4	5
Prob(C)	0.8	0.6	–	–	–
Prob(G)	0.2	0.4	0.8	–	–
Prob(H)	–	–	0.2	–	–
Prob(S)	–	–	–	0.6	0.2
Prob(T)	–	–	–	0.4	–
Prob(L)	–	–	–	–	0.6
Prob(V)	–	–	–	–	0.2

• Given this model the probability of CGGSV is:
0.8 * 0.4 * 0.8* 0.6* 0.2 = 0.031

• Since multiplication of fractions is computationally expensive and prone to floating point errors, a transformation into the logarithmic world is used.

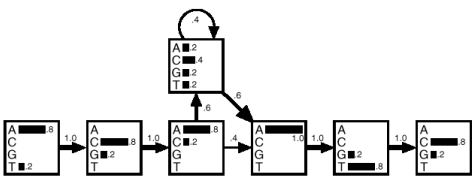
• The score is calculated by taking the logs of all amino acid probabilities and adding them up.
ln(0.8) + ln(0.4) + ln(0.8) + ln(0.6) + ln(0.2) = -3.48

Stochastic modeling of biological sequences

A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C

[AT] [CG] [AC] [ACGT]* A [TG] [GC]

But with this expression it is not possible to distinguish between the highly implausible sequence TGCT- - AGG and the consensus sequence ACAC - - ATC



The HMM architecture

- S-start; E-end
- m- main state (matches/mismatches)
- i - insert state
- d - delete state

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

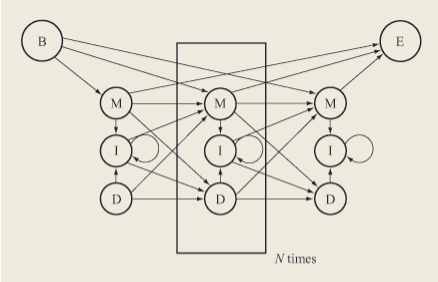


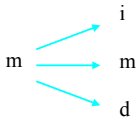
Figure 3

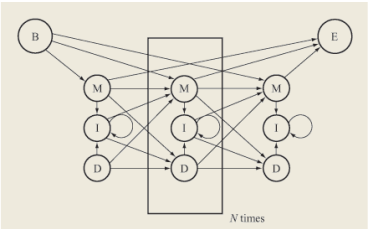
A profile HMM, which has a repetitive structure of three states (M, I, and D). Each set of three states represents a single column in the alignment of protein sequences.

Parameters used in HMM building

- Transition probability: T_{ij} (average 0.333)
- Emission probability: $E_{i\alpha}$ (average 0.05)

```
M N - F L S
M N - F L S
M N K Y L T
M Q - W - T
```

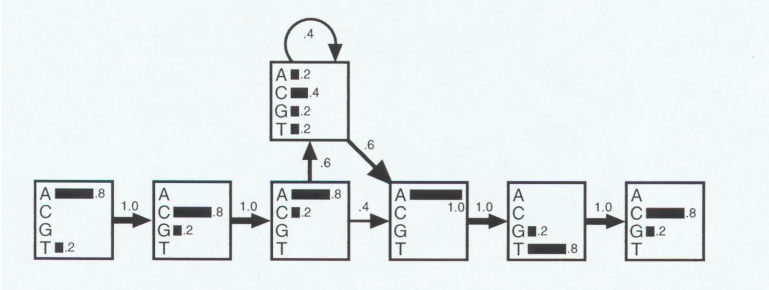




- Since the probabilities are very small numbers, they are converted to log odds scores and added to get the overall probability score

Markov modeling of biological sequences

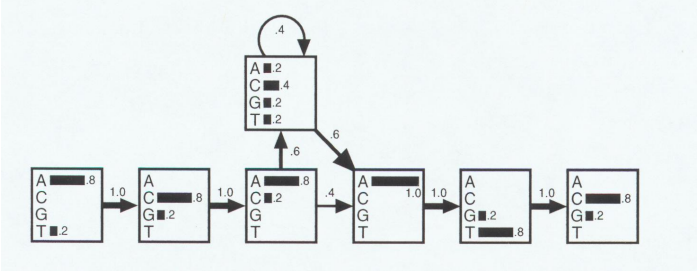
```
1) A C A - - - A T G
2) T C A A C T A T C
3) A C A C - - A G C
4) A G A - - - A T C
5) A C C G - - A T C
```



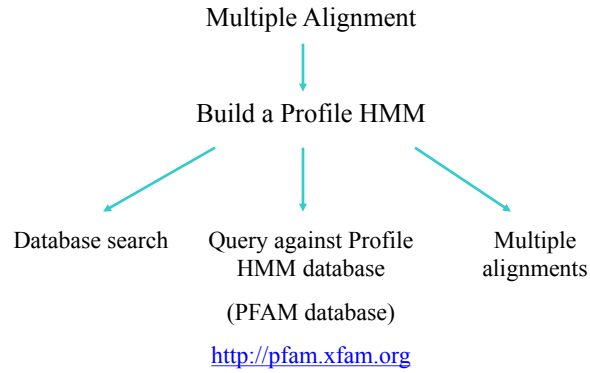
Markov modeling of biological sequences

	P(s)*100
1. A C A - - - A T G	3.3
2. T C A A C T A T C	0.0075
3. A C A C - - A G C	1.2
4. A G A - - - A T C	3.3
5. A C C G - - A T C	0.59
A C A C - - A T C	4.7

P(ACACATC)= 0.047 Obtained by taking the product of probabilities for residues in each state and the transitions.



Sequence Alignment and Database Search using HMMER



(PFAM database)

http://pfam.xfam.org

PFAM http://pfam.xfam.org

- Protein Family Database created using HMMs
- Pfam-A contains functionally annotated families (16,712 families)
  - About 75% sequence coverage and 52% residue coverage
  - DUFs (Domain of unknown functions)
- Pfam-B contains unannotated families (Over 500,000)
- All protein sequences were clustered into families based on sequence identity
- Seed multiple alignments were built using ClustalW and manual checking
- HMM models were built using the HMMER suite
- Using these models, more family members are added in an iterative process to update the HMM Models until no more new members are found

Other HMM-based resources

- Rfams: A collection of HMM models based on RNA families obtained by multiple sequence alignments, consensus 2-D structure and covariance models.
  - http://rfam.xfam.org
- Dfams: A collection of HMM models based on repetitive DNA sequence elements
  - http://www.dfam.org
- Antifams: Collection of Profile-HMMs built from commonly occurring non-coding RNAs (such as tRNAs) which can be used as a quality control to identify spurious protein predictions.
  - ftp://ftp.sanger.ac.uk/pub/databases/Pfam/AntiFam/