Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 18

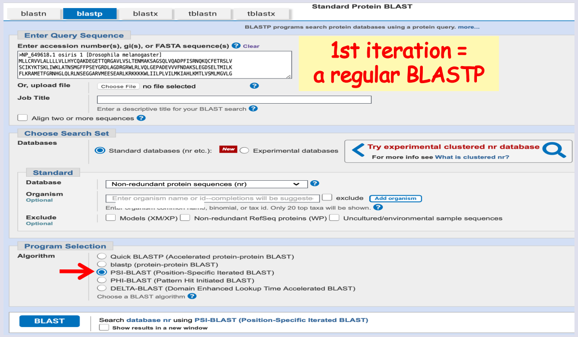
1

TODAY'S TOPICS

- ➢ PSI-BLAST
- ➢ Hidden Markov model and profile HMM
- ➢ Protein family/domain databases
  (InterPro, HMMER, etc.)

2

PSI-BLAST



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

```
>NP_649618.1 osiris 1 [Drosophila melanogaster]
MLLCRVVLALLLLVLLHYCQAKDEGETTQRGAVLVSLTENMAKSAGSQLVQADPFISRNQKQCFETRSLV
SCIKYKTSKLIWKLATNSMGFFPSEYGRDLAGDRGRWLRLVQLGEPADEVVVFNDAKSLEGDSELTMILK
FLKRAMETFGRNHGLQLRLNSEGGARVMEESEARLKRKKKKWLIILPLVILMKIAHLKMTLVSMLMGVLG
```

Or, upload file: Choose File no file selected ?

Job Title: Enter a descriptive title for your BLAST search ?

□ Align two or more sequences ?

Choose Search Set

Databases: ● Standard databases (nr etc.): New ○ Experimental databases

Try experimental clustered nr database. For more info see What is clustered nr?

Standard

Database: Non-redundant protein sequences (nr) ?

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. ?

Exclude (Optional): □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

Algorithm:
- ○ Quick BLASTP (Accelerated protein-protein BLAST)
- ○ blastp (protein-protein BLAST)
- ● PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)
- ○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm ?

BLAST — Search database nr using PSI-BLAST (Position-Specific Iterated BLAST)

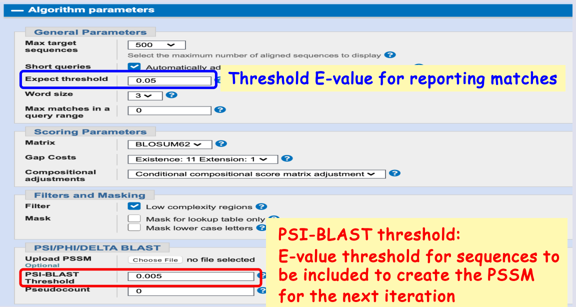
□ Show results in a new window

1st iteration =
a regular BLASTP

3

PSI-BLAST

- ➢ Two different E-value thresholds



— Algorithm parameters

General Parameters

Max target sequences: 500

Select the maximum number of aligned sequences to display ?

Short queries: ☑ Automatically ad

Expect threshold: 0.05

Word size: 3 ?

Max matches in a query range: 0 ?

Scoring Parameters

Matrix: BLOSUM62 ?

Gap Costs: Existence: 11 Extension: 1 ?

Compositional adjustments: Conditional compositional score matrix adjustment ?

Filters and Masking

Filter: ☑ Low complexity regions ?

Mask: □ Mask for lookup table only □ Mask lower case letters ?

PSI/PHI/DELTA BLAST

Upload PSSM (Optional): Choose File no file selected

PSI-BLAST Threshold: 0.005

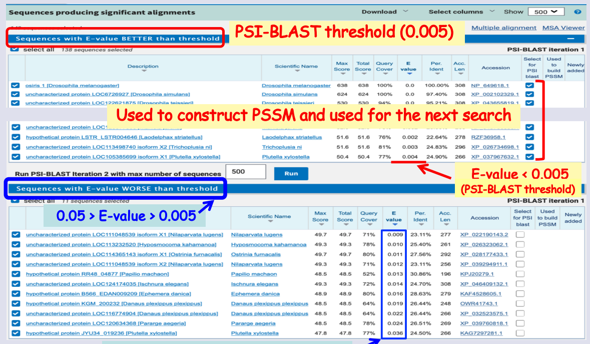
Pseudocount: 0

Threshold E-value for reporting matches

PSI-BLAST threshold:
E-value threshold for sequences to be included to create the PSSM for the next iteration

4

PSI-BLAST



Sequences producing significant alignments — Download — Select columns — Show 500 ?

Multiple alignment MSA Viewer

Sequences with E-value BETTER than threshold

☑ select all 138 sequences selected — PSI-BLAST iteration 1

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
☑	osiris 1 [Drosophila melanogaster]	Drosophila melanogaster	638	638	100%	0.0	100.00%	308	NP_649618.1	☑		
☑	uncharacterized protein LOC6726927 [Drosophila simulans]	Drosophila simulans	624	624	100%	0.0	97.40%	308	XP_002102329.1	☑		
☑	uncharacterized protein LOC122621875 [Drosophila teissieri]	Drosophila teissieri	530	530	94%	0.0	95.21%	308	XP_043655819.1	☑		
☑	uncharacterized protein LOC1									☑		
☑	hypothetical protein LSTR_LSTR004646 [Laodelphax striatellus]	Laodelphax striatellus	51.6	51.6	76%	0.002	22.64%	278	RZF36958.1	☑		
☑	uncharacterized protein LOC113498740 isoform X2 [Trichoplusia ni]	Trichoplusia ni	51.6	51.6	81%	0.003	24.83%	296	XP_026734698.1	☑		
☑	uncharacterized protein LOC105385699 isoform X1 [Plutella xylostella]	Plutella xylostella	50.4	50.4	77%	0.004	24.90%	266	XP_037967632.1	☑		

Run PSI-BLAST Iteration 2 with max number of sequences 500 Run

Sequences with E-value WORSE than threshold

☑ select all 11 sequences selected — PSI-BLAST iteration 1

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
☑	uncharacterized protein LOC111048539 isoform X1 [Nilaparvata lugens]	Nilaparvata lugens	49.7	49.7	71%	0.009	23.11%	277	XP_022190143.2	□		
☑	uncharacterized protein LOC113232520 [Hyposmocoma kahamanoa]	Hyposmocoma kahamanoa	49.3	49.3	78%	0.010	25.40%	261	XP_026323062.1	□		
☑	uncharacterized protein LOC114365143 isoform X1 [Ostrinia furnacalis]	Ostrinia furnacalis	49.7	49.7	80%	0.011	27.56%	292	XP_028177433.1	□		
☑	uncharacterized protein LOC111048539 isoform X2 [Nilaparvata lugens]	Nilaparvata lugens	49.3	49.3	71%	0.012	23.11%	256	XP_039294911.1	□		
☑	hypothetical protein RR48_04877 [Papilio machaon]	Papilio machaon	48.5	48.5	52%	0.013	30.86%	196	KPJ20279.1	□		
☑	uncharacterized protein LOC124174035 [Ischnura elegans]	Ischnura elegans	49.3	49.3	72%	0.014	24.70%	308	XP_046409132.1	□		
☑	hypothetical protein B566_EDAN009209 [Ephemera danica]	Ephemera danica	48.9	48.9	80%	0.016	28.63%	279	KAF4528605.1	□		
☑	hypothetical protein KGM_200232 [Danaus plexippus plexippus]	Danaus plexippus plexippus	48.5	48.5	64%	0.019	26.44%	248	OWR41743.1	□		
☑	uncharacterized protein LOC116774904 [Danaus plexippus plexippus]	Danaus plexippus plexippus	48.5	48.5	64%	0.022	26.44%	266	XP_032523575.1	□		
☑	uncharacterized protein LOC120634368 [Pararge aegeria]	Pararge aegeria	48.5	48.5	78%	0.024	26.51%	269	XP_039760818.1	□		
☑	hypothetical protein JYU34_019236 [Plutella xylostella]	Plutella xylostella	47.8	47.8	77%	0.036	24.50%	266	KAG7297281.1	□		

PSI-BLAST threshold (0.005)

Used to construct PSSM and used for the next search

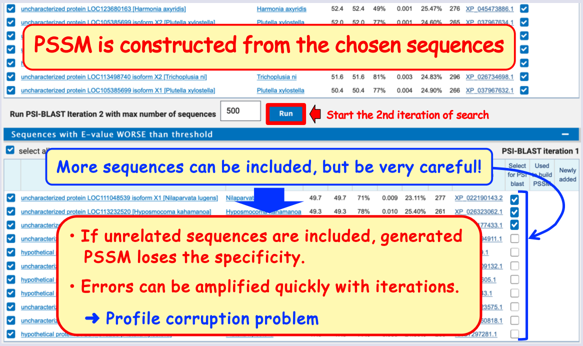
E-value < 0.005
(PSI-BLAST threshold)

0.05 > E-value > 0.005

E-value < 0.05 (reporting threshold)

5

PSI-BLAST



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast
☑	uncharacterized protein LOC123680163 [Harmonia axyridis]	Harmonia axyridis	52.4	52.4	49%	0.001	25.47%	276	XP_045473886.1	☑
☑	uncharacterized protein LOC105385699 isoform X2 [Plutella xylostella]	Plutella xylostella	52.0	52.0	77%	0.001	24.60%	265	XP_037967634.1	☑
☑										☑
☑										☑
☑										☑
☑	uncharacterized protein LOC113498740 isoform X2 [Trichoplusia ni]	Trichoplusia ni	51.6	51.6	81%	0.003	24.83%	296	XP_026734698.1	☑
☑	uncharacterized protein LOC105385699 isoform X1 [Plutella xylostella]	Plutella xylostella	50.4	50.4	77%	0.004	24.90%	266	XP_037967632.1	☑

Run PSI-BLAST Iteration 2 with max number of sequences 500 Run

Sequences with E-value WORSE than threshold

☑ select a — PSI-BLAST iteration 1

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
☑	uncharacterized protein LOC111048539 isoform X1 [Nilaparvata lugens]	Nilaparvat	49.7	49.7	71%	0.009	23.11%	277	XP_022190143.2	☑		
☑	uncharacterized protein LOC113232520 [Hyposmocoma kahamanoa]	Hyposmocoma kahamanoa	49.3	49.3	78%	0.010	25.40%	261	XP_026323062.1	☑		
☑	uncharacteriz								77433.1	☑		
☑	uncharacteri								94911.1	□		
☑	hypothetical								.1	□		
☑	uncharacteri								09132.1	□		
☑	hypothetical								605.1	□		
☑	hypothetical								43.1	□		
☑	uncharacteri								23575.1	□		
☑	uncharacteri								60818.1	□		
☑	hypothetical prot								297281.1	□		

PSSM is constructed from the chosen sequences

Start the 2nd iteration of search

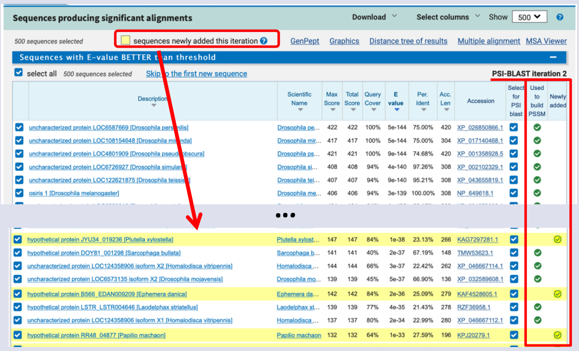
More sequences can be included, but be very careful!

- If unrelated sequences are included, generated PSSM loses the specificity.
- Errors can be amplified quickly with iterations.

  ➜ Profile corruption problem

6

PSI-BLAST



Sequences producing significant alignments — Download — Select columns — Show 500

500 sequences selected — □ sequences newly added this iteration — GenPept — Graphics — Distance tree of results — Multiple alignment — MSA Viewer

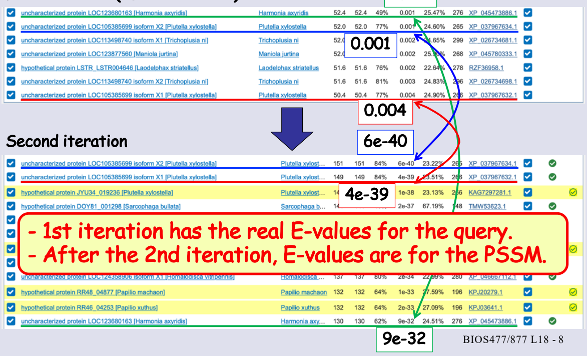
Sequences with E-value BETTER than threshold

☑ select all — 500 sequences selected — Skip to the first new sequence — PSI-BLAST iteration 2

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
☑	uncharacterized protein LOC6587669 [Drosophila persimilis]	Drosophila pe...	422	422	100%	5e-144	75.00%	420	XP_026850866.1	☑	✓	
☑	uncharacterized protein LOC108154648 [Drosophila miranda]	Drosophila mir...	417	417	100%	5e-144	75.00%	304	XP_017140468.1	☑	✓	
☑	uncharacterized protein LOC4801909 [Drosophila pseudoobscura]	Drosophila ps...	421	421	100%	9e-144	74.68%	420	XP_001358928.5	☑	✓	
☑	uncharacterized protein LOC6726927 [Drosophila simulans]	Drosophila si...	408	408	94%	4e-140	97.26%	308	XP_002102329.1	☑	✓	
☑	uncharacterized protein LOC122621875 [Drosophila teissieri]	Drosophila tei...	407	407	94%	9e-140	95.21%	308	XP_043655819.1	☑	✓	
☑	osiris 1 [Drosophila melanogaster]	Drosophila me...	406	406	94%	3e-139	100.00%	308	NP_649618.1	☑	✓	
•••												
☑	hypothetical protein JYU34_019236 [Plutella xylostella]	Plutella xylost...	147	147	84%	1e-38	23.13%	266	KAG7297281.1	☑		✓
☑	hypothetical protein DOY81_001298 [Sarcophaga bullata]	Sarcophaga b...	141	141	40%	2e-37	67.19%	148	TMW53623.1	☑	✓	
☑	uncharacterized protein LOC124358906 isoform X2 [Homalodisca vitripennis]	Homalodisca ...	144	144	66%	3e-37	22.42%	262	XP_046667114.1	☑	✓	
☑	uncharacterized protein LOC6573135 isoform X2 [Drosophila mojavensis]	Drosophila mo...	139	139	45%	5e-37	66.90%	136	XP_032589608.1	☑	✓	
☑	hypothetical protein B566_EDAN009209 [Ephemera danica]	Ephemera da...	142	142	84%	2e-36	25.09%	279	KAF4528605.1	☑		✓
☑	hypothetical protein LSTR_LSTR004646 [Laodelphax striatellus]	Laodelphax st...	139	139	77%	4e-35	21.43%	278	RZF36958.1	☑	✓	
☑	uncharacterized protein LOC124358906 isoform X1 [Homalodisca vitripennis]	Homalodisca ...	137	137	80%	2e-34	22.99%	280	XP_046667112.1	☑	✓	
☑	hypothetical protein RR48_04877 [Papilio machaon]	Papilio machaon	132	132	64%	1e-33	27.59%	196	KPJ20279.1	☑		✓

7

PSI-BLAST



First iteration (BLASTP search)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast
☑	uncharacterized protein LOC123680163 [Harmonia axyridis]	Harmonia axyridis	52.4	52.4	49%	0.001	25.47%	276	XP_045473886.1	☑
☑	uncharacterized protein LOC105385699 isoform X2 [Plutella xylostella]	Plutella xylostella	52.0	52.0	77%	0.001	24.60%	265	XP_037967634.1	☑
☑	uncharacterized protein LOC113498740 isoform X1 [Trichoplusia ni]	Trichoplusia ni	52.0			0.002	24.65%	299	XP_026734681.1	☑
☑	uncharacterized protein LOC123877560 [Maniola jurtina]	Maniola jurtina	52.0			0.002	25.30%	268	XP_045780333.1	☑
☑	hypothetical protein LSTR_LSTR004646 [Laodelphax striatellus]	Laodelphax striatellus	51.6	51.6	76%	0.002	22.64%	278	RZF36958.1	☑
☑	uncharacterized protein LOC113498740 isoform X2 [Trichoplusia ni]	Trichoplusia ni	51.6	51.6	81%	0.003	24.83%	296	XP_026734698.1	☑
☑	uncharacterized protein LOC105385699 isoform X1 [Plutella xylostella]	Plutella xylostella	50.4	50.4	77%	0.004	24.90%	266	XP_037967632.1	☑

0.001

0.001

0.004

Second iteration

6e-40

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
☑	uncharacterized protein LOC105385699 isoform X2 [Plutella xylostella]	Plutella xylost...	151	151	84%	6e-40	23.22%	265	XP_037967634.1	☑	✓	
☑	uncharacterized protein LOC105385699 isoform X1 [Plutella xylostella]	Plutella xylost...	149	149	84%	4e-39	23.51%	266	XP_037967632.1	☑	✓	
☑	hypothetical protein JYU34_019236 [Plutella xylostella]	Plutella xylost...	14			1e-38	23.13%	266	KAG7297281.1	☑		✓
☑	hypothetical protein DOY81_001298 [Sarcophaga bullata]	Sarcophaga b...	14			2e-37	67.19%	148	TMW53623.1	☑	✓	
☑	uncharacterized protein LOC124358906 isoform X1 [Homalodisca vitripennis]	Homalodisca ...	137	137	80%	2e-34	22.99%	280	XP_046667112.1	☑	✓	
☑	hypothetical protein RR48_04877 [Papilio machaon]	Papilio machaon	132	132	64%	1e-33	27.59%	196	KPJ20279.1	☑		✓
☑	hypothetical protein RR46_04253 [Papilio xuthus]	Papilio xuthus	132	132	64%	2e-33	27.09%	196	KPJ03641.1	☑		✓
☑	uncharacterized protein LOC123680163 [Harmonia axyridis]	Harmonia axy...	130	130	62%	9e-32	24.51%	276	XP_045473886.1	☑	✓	

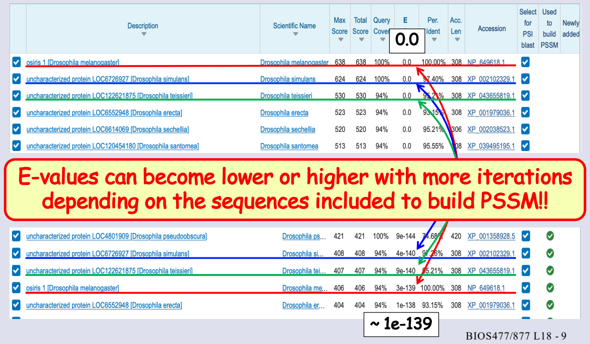
4e-39

- 1st iteration has the real E-values for the query.
- After the 2nd iteration, E-values are for the PSSM.

9e-32

8

PSI-BLAST



First iteration (BLASTP search)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
☑	osiris 1 [Drosophila melanogaster]	Drosophila melanogaster	638	638	100%	0.0	100.00%	308	NP_649618.1	☑		
☑	uncharacterized protein LOC6726927 [Drosophila simulans]	Drosophila simulans	624	624	100%	0.0	97.40%	308	XP_002102329.1	☑		
☑	uncharacterized protein LOC122621875 [Drosophila teissieri]	Drosophila teissieri	530	530	94%	0.0	95.21%	308	XP_043655819.1	☑		
☑	uncharacterized protein LOC6552948 [Drosophila erecta]	Drosophila erecta	523	523	94%	0.0	93.15%	308	XP_001979036.1	☑		
☑	uncharacterized protein LOC6614069 [Drosophila sechellia]	Drosophila sechellia	520	520	94%	0.0	95.21%	306	XP_002038523.1	☑		
☑	uncharacterized protein LOC120454180 [Drosophila santomea]	Drosophila santomea	513	513	94%	0.0	95.55%	308	XP_039495195.1	☑		

0.0

E-values can become lower or higher with more iterations depending on the sequences included to build PSSM!!

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM
☑	uncharacterized protein LOC4801909 [Drosophila pseudoobscura]	Drosophila ps...	421	421	100%	9e-144	74.68%	420	XP_001358928.5	☑	✓
☑	uncharacterized protein LOC6726927 [Drosophila simulans]	Drosophila si...	408	408	94%	4e-140	97.26%	308	XP_002102329.1	☑	✓
☑	uncharacterized protein LOC122621875 [Drosophila teissieri]	Drosophila tei...	407	407	94%	9e-140	95.21%	308	XP_043655819.1	☑	✓
☑	osiris 1 [Drosophila melanogaster]	Drosophila me...	406	406	94%	3e-139	100.00%	308	NP_649618.1	☑	✓
☑	uncharacterized protein LOC6552948 [Drosophila erecta]	Drosophila er...	404	404	94%	1e-138	93.15%	308	XP_001979036.1	☑	✓

~ 1e-139

9

Profile hidden Markov models

- ➢ **Probabilistic models of multiple alignments**

  (Krogh *et al.*, 1994; Krogh 1998)

  ➜ Closely related to standard profiles (PSSMs) introduced by Gribskov

  ➜ Used in *e.g.,* PFAM, SMART, Superfamily, Panther (databases of multiple alignments and profile HMMs)

What is a hidden Markov model?

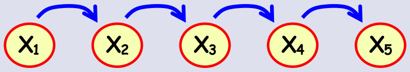
What is hidden?

Eddy (2004) What is a hidden Markov model? ← DNA regulatory element example
Nature Biotechnology 22: 1315-1316.

10

Markov models

- ➢ Markov chain: a stochastic process from one state to the next that exhibits the Markov property (*e.g.*, States for DNA: A, T, G, C)



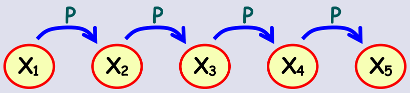
- Markov property: The next state of the system depends only on the present state of the system [No memory]

11

Markov models

- ➢ Markov chain: a stochastic process from one state to the next that exhibits the Markov property (*e.g.*, States for DNA: A, T, G, C)

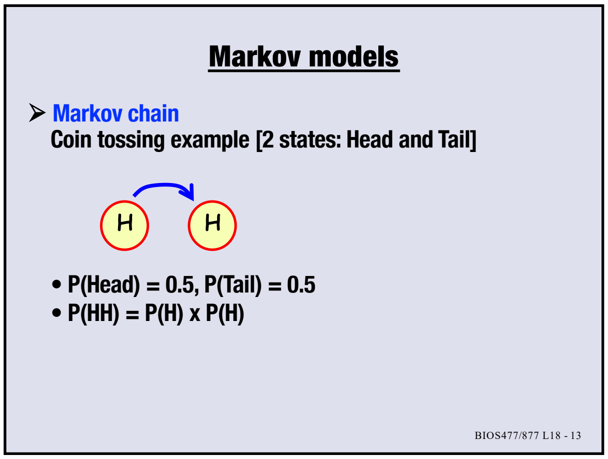
[P: Transition probability]



- Markov property: The next state of the system depends only on the present state of the system [No memory]
- Transition probabilities are independent of time [Time homogeneous]

- Markov chain if the state spaces is discrete
- Markov process if the state space is continuous

12



Markov models

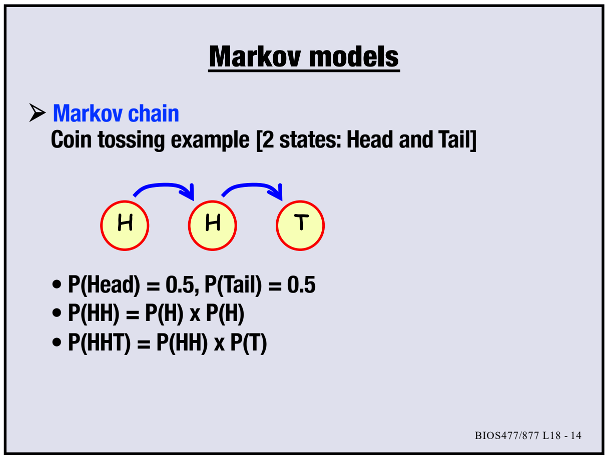
- Markov chain

  Coin tossing example [2 states: Head and Tail]

- P(Head) = 0.5, P(Tail) = 0.5
- P(HH) = P(H) x P(H)

BIOS477/877 L18 - 13

13



Markov models

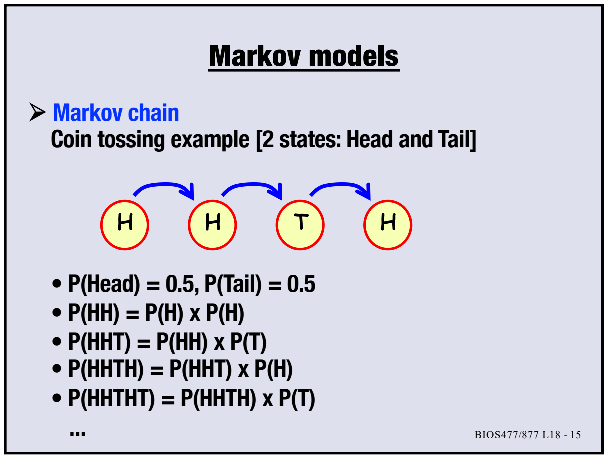
- Markov chain

  Coin tossing example [2 states: Head and Tail]

- P(Head) = 0.5, P(Tail) = 0.5
- P(HH) = P(H) x P(H)
- P(HHT) = P(HH) x P(T)

BIOS477/877 L18 - 14

14



Markov models

- Markov chain

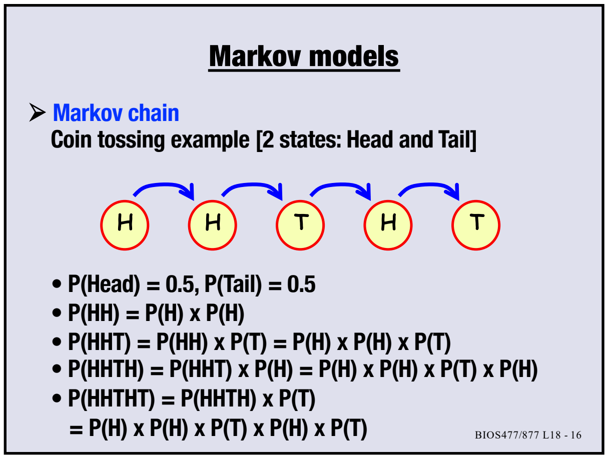
  Coin tossing example [2 states: Head and Tail]

- P(Head) = 0.5, P(Tail) = 0.5
- P(HH) = P(H) x P(H)
- P(HHT) = P(HH) x P(T)
- P(HHTH) = P(HHT) x P(H)
- P(HHTHT) = P(HHTH) x P(T)

  …

BIOS477/877 L18 - 15

15



Markov models

- Markov chain

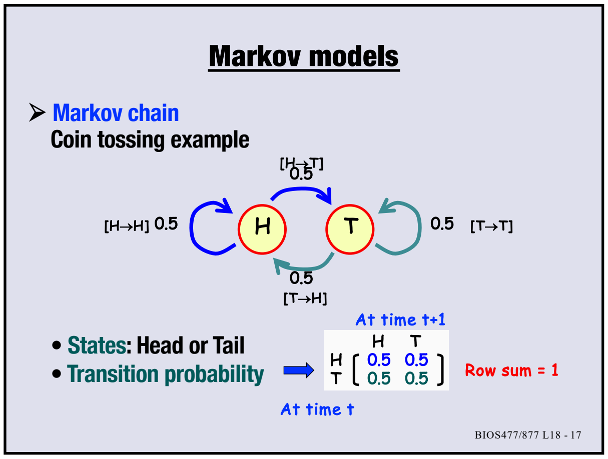
  Coin tossing example [2 states: Head and Tail]

- P(Head) = 0.5, P(Tail) = 0.5
- P(HH) = P(H) x P(H)
- P(HHT) = P(HH) x P(T) = P(H) x P(H) x P(T)
- P(HHTH) = P(HHT) x P(H) = P(H) x P(H) x P(T) x P(H)
- P(HHTHT) = P(HHTH) x P(T)

  = P(H) x P(H) x P(T) x P(H) x P(T)

BIOS477/877 L18 - 16

16



Markov models

- Markov chain

  Coin tossing example

- States: Head or Tail
- Transition probability ➡

At time t+1

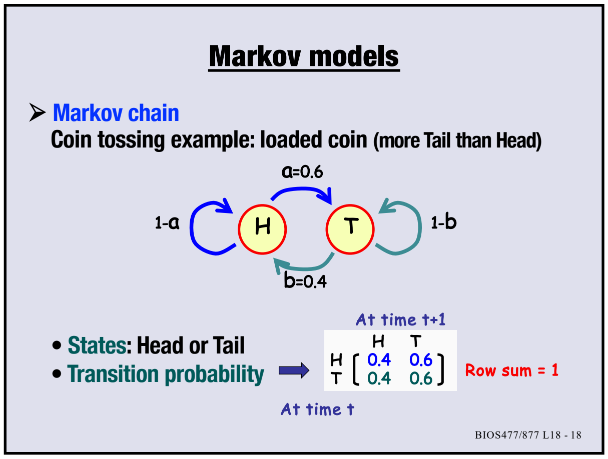
$$\begin{array}{c|cc} & H & T \\ \hline H & 0.5 & 0.5 \\ T & 0.5 & 0.5 \end{array}$$

Row sum = 1

At time t

BIOS477/877 L18 - 17

17



Markov models

- Markov chain

  Coin tossing example: loaded coin (more Tail than Head)

- States: Head or Tail
- Transition probability ➡

At time t+1

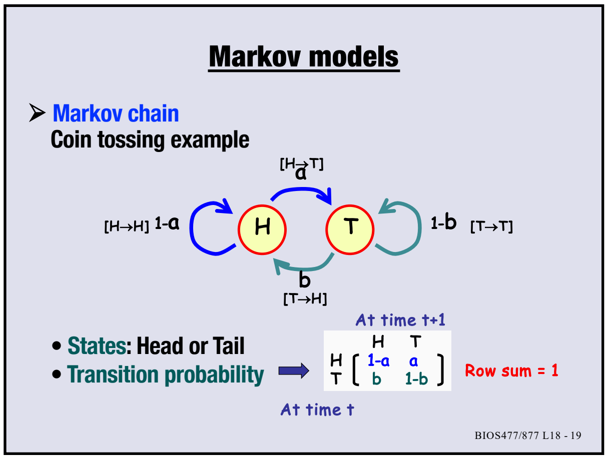
$$\begin{array}{c|cc} & H & T \\ \hline H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

Row sum = 1

At time t

BIOS477/877 L18 - 18

18



Markov models

- Markov chain

Coin tossing example

- States: Head or Tail
- Transition probability

At time t+1

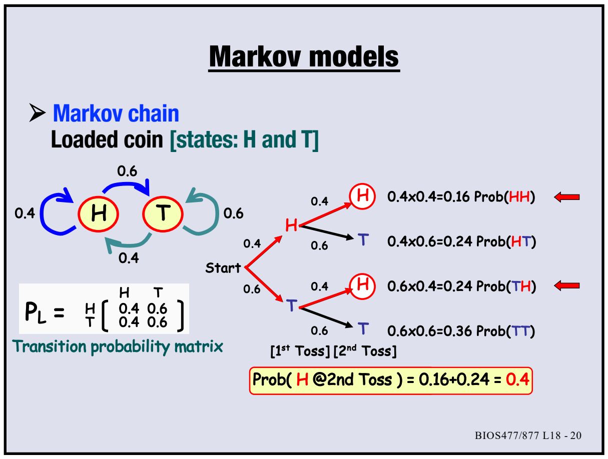
$$\begin{array}{ccc} & H & T \\ H & 1-a & a \\ T & b & 1-b \end{array}$$

At time t

Row sum = 1

BIOS477/877 L18 - 19

19



Markov models

- Markov chain

Loaded coin [states: H and T]

0.4x0.4=0.16 Prob(HH)

0.4x0.6=0.24 Prob(HT)

0.6x0.4=0.24 Prob(TH)

0.6x0.6=0.36 Prob(TT)

[1st Toss] [2nd Toss]

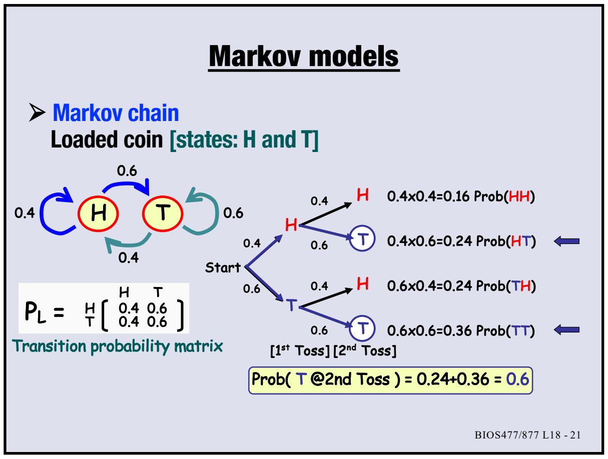
$$P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

Transition probability matrix

Prob(H @2nd Toss) = 0.16+0.24 = 0.4

BIOS477/877 L18 - 20

20



Markov models

- Markov chain

Loaded coin [states: H and T]

0.4x0.4=0.16 Prob(HH)

0.4x0.6=0.24 Prob(HT)

0.6x0.4=0.24 Prob(TH)

0.6x0.6=0.36 Prob(TT)

[1st Toss] [2nd Toss]

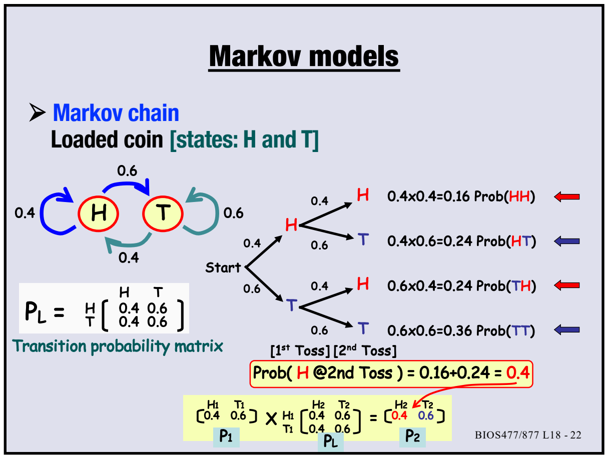
$$P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

Transition probability matrix

Prob(T @2nd Toss) = 0.24+0.36 = 0.6

BIOS477/877 L18 - 21

21



Markov models

- Markov chain

Loaded coin [states: H and T]

0.4x0.4=0.16 Prob(HH)

0.4x0.6=0.24 Prob(HT)

0.6x0.4=0.24 Prob(TH)

0.6x0.6=0.36 Prob(TT)

[1st Toss] [2nd Toss]

$$P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

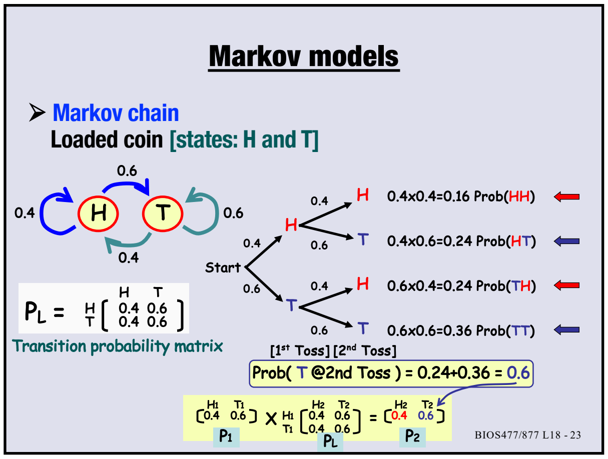
Transition probability matrix

Prob(H @2nd Toss) = 0.16+0.24 = 0.4

$$\underbrace{\begin{array}{cc} H_1 & T_1 \\ 0.4 & 0.6 \end{array}}_{P_1} \times \underbrace{\begin{array}{ccc} & H_2 & T_2 \\ H_1 & 0.4 & 0.6 \\ T_1 & 0.4 & 0.6 \end{array}}_{P_L} = \underbrace{\begin{array}{cc} H_2 & T_2 \\ 0.4 & 0.6 \end{array}}_{P_2}$$

BIOS477/877 L18 - 22

22



Markov models

- Markov chain

Loaded coin [states: H and T]

0.4x0.4=0.16 Prob(HH)

0.4x0.6=0.24 Prob(HT)

0.6x0.4=0.24 Prob(TH)

0.6x0.6=0.36 Prob(TT)

[1st Toss] [2nd Toss]

$$P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

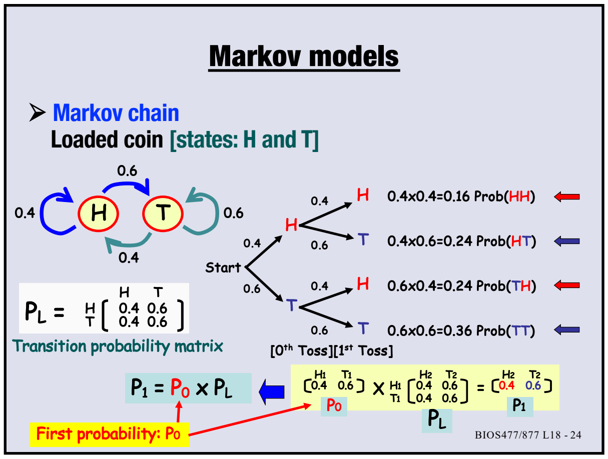
Transition probability matrix

Prob(T @2nd Toss) = 0.24+0.36 = 0.6

$$\underbrace{\begin{array}{cc} H_1 & T_1 \\ 0.4 & 0.6 \end{array}}_{P_1} \times \underbrace{\begin{array}{ccc} & H_2 & T_2 \\ H_1 & 0.4 & 0.6 \\ T_1 & 0.4 & 0.6 \end{array}}_{P_L} = \underbrace{\begin{array}{cc} H_2 & T_2 \\ 0.4 & 0.6 \end{array}}_{P_2}$$

BIOS477/877 L18 - 23

23



Markov models

- Markov chain

Loaded coin [states: H and T]

0.4x0.4=0.16 Prob(HH)

0.4x0.6=0.24 Prob(HT)

0.6x0.4=0.24 Prob(TH)

0.6x0.6=0.36 Prob(TT)

[0th Toss][1st Toss]

$$P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

Transition probability matrix

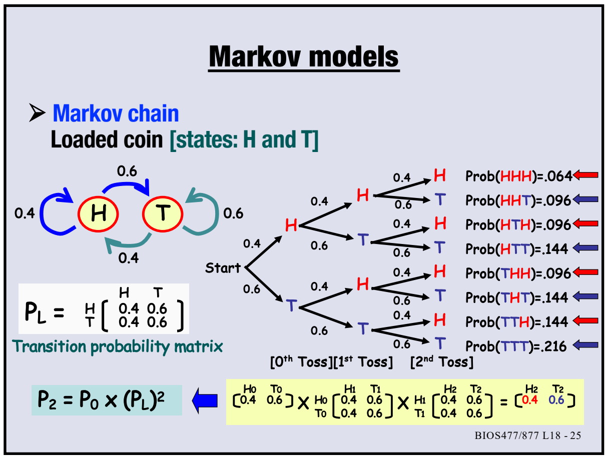
$$P_1 = P_0 \times P_L$$

$$\underbrace{\begin{array}{cc} H_1 & T_1 \\ 0.4 & 0.6 \end{array}}_{P_0} \times \underbrace{\begin{array}{ccc} & H_2 & T_2 \\ H_1 & 0.4 & 0.6 \\ T_1 & 0.4 & 0.6 \end{array}}_{P_L} = \underbrace{\begin{array}{cc} H_2 & T_2 \\ 0.4 & 0.6 \end{array}}_{P_1}$$

First probability: P_0

BIOS477/877 L18 - 24

24



Markov models

- **Markov chain**
  Loaded coin [states: H and T]

Prob(HHH)=.064
Prob(HHT)=.096
Prob(HTH)=.096
Prob(HTT)=.144
Prob(THH)=.096
Prob(THT)=.144
Prob(TTH)=.144
Prob(TTT)=.216

[0th Toss][1st Toss] [2nd Toss]

$$P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

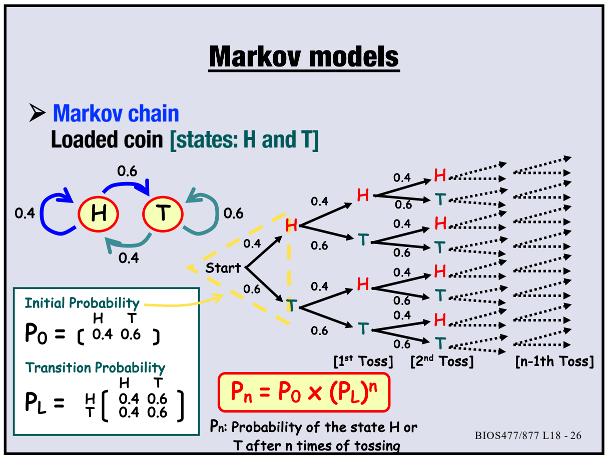
Transition probability matrix

$$P_2 = P_0 \times (P_L)^2$$

$$\begin{array}{cc} H_0 & T_0 \\ 0.4 & 0.6 \end{array} \times \begin{array}{ccc} & H_1 & T_1 \\ H_0 & 0.4 & 0.6 \\ T_0 & 0.4 & 0.6 \end{array} \times \begin{array}{ccc} & H_2 & T_2 \\ H_1 & 0.4 & 0.6 \\ T_1 & 0.4 & 0.6 \end{array} = \begin{array}{cc} H_2 & T_2 \\ 0.4 & 0.6 \end{array}$$

BIOS477/877 L18 - 25

25



Markov models

- **Markov chain**
  Loaded coin [states: H and T]

Initial Probability

$$P_0 = \begin{array}{cc} H & T \\ 0.4 & 0.6 \end{array}$$

Transition Probability

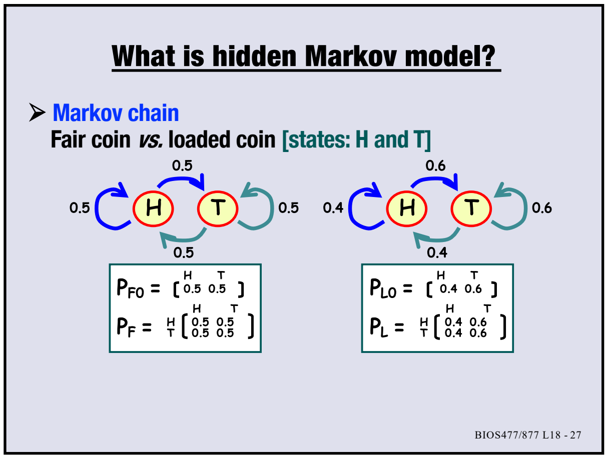
$$P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

$$P_n = P_0 \times (P_L)^n$$

P_n: Probability of the state H or T after n times of tossing

BIOS477/877 L18 - 26

26



What is hidden Markov model?

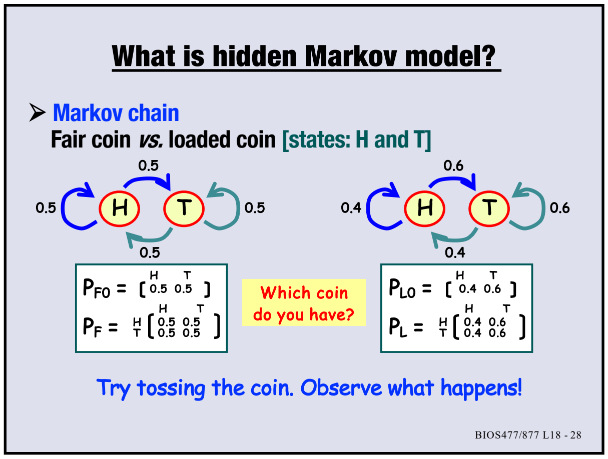
- **Markov chain**
  Fair coin *vs.* loaded coin [states: H and T]

$$P_{F0} = \begin{array}{cc} H & T \\ 0.5 & 0.5 \end{array} \qquad P_F = \begin{array}{ccc} & H & T \\ H & 0.5 & 0.5 \\ T & 0.5 & 0.5 \end{array}$$

$$P_{L0} = \begin{array}{cc} H & T \\ 0.4 & 0.6 \end{array} \qquad P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

BIOS477/877 L18 - 27

27



What is hidden Markov model?

- **Markov chain**
  Fair coin *vs.* loaded coin [states: H and T]

$$P_{F0} = \begin{array}{cc} H & T \\ 0.5 & 0.5 \end{array} \qquad P_F = \begin{array}{ccc} & H & T \\ H & 0.5 & 0.5 \\ T & 0.5 & 0.5 \end{array}$$

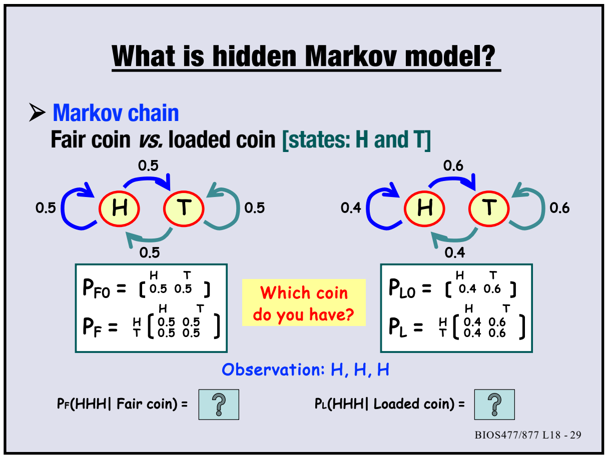
Which coin do you have?

$$P_{L0} = \begin{array}{cc} H & T \\ 0.4 & 0.6 \end{array} \qquad P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

Try tossing the coin. Observe what happens!

BIOS477/877 L18 - 28

28



What is hidden Markov model?

- **Markov chain**
  Fair coin *vs.* loaded coin [states: H and T]

$$P_{F0} = \begin{array}{cc} H & T \\ 0.5 & 0.5 \end{array} \qquad P_F = \begin{array}{ccc} & H & T \\ H & 0.5 & 0.5 \\ T & 0.5 & 0.5 \end{array}$$

Which coin do you have?

$$P_{L0} = \begin{array}{cc} H & T \\ 0.4 & 0.6 \end{array} \qquad P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

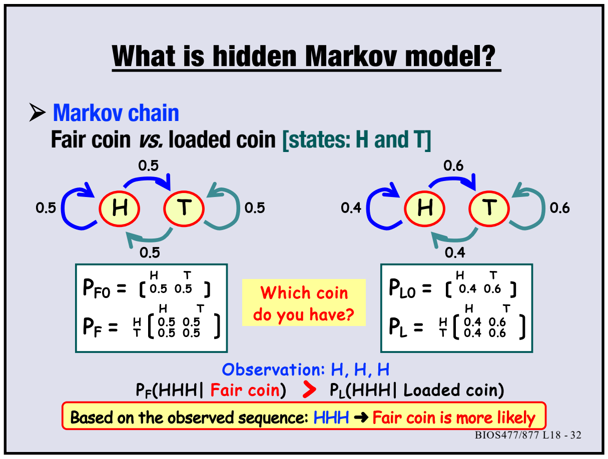
Observation: H, H, H

P_F(HHH| Fair coin) = ?

P_L(HHH| Loaded coin) = ?

BIOS477/877 L18 - 29

29



What is hidden Markov model?

- **Markov chain**
  Fair coin *vs.* loaded coin [states: H and T]

$$P_{F0} = \begin{array}{cc} H & T \\ 0.5 & 0.5 \end{array} \qquad P_F = \begin{array}{ccc} & H & T \\ H & 0.5 & 0.5 \\ T & 0.5 & 0.5 \end{array}$$

Which coin do you have?

$$P_{L0} = \begin{array}{cc} H & T \\ 0.4 & 0.6 \end{array} \qquad P_L = \begin{array}{ccc} & H & T \\ H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$$

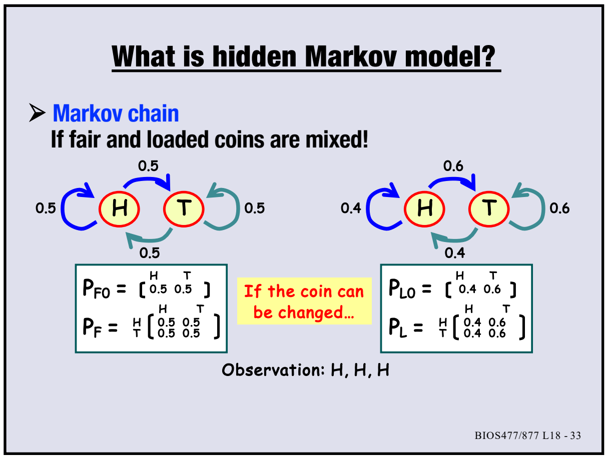
Observation: H, H, H

P_F(HHH| Fair coin) > P_L(HHH| Loaded coin)

Based on the observed sequence: HHH → Fair coin is more likely

BIOS477/877 L18 - 32

32



What is hidden Markov model?

- **Markov chain**
  If fair and loaded coins are mixed!

0.5
0.5
H
T
0.5
0.5
0.6
0.4
H
T
0.6
0.4

$P_{F0} = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ [\,0.5 \;\; 0.5\,] \end{array}$

$P_F = \begin{array}{c|cc} & H & T \\ \hline H & 0.5 & 0.5 \\ T & 0.5 & 0.5 \end{array}$

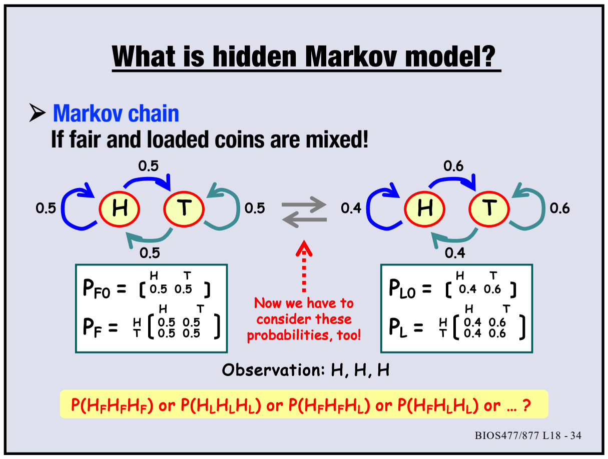
If the coin can be changed…

$P_{L0} = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ [\,0.4 \;\; 0.6\,] \end{array}$

$P_L = \begin{array}{c|cc} & H & T \\ \hline H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$

Observation: H, H, H

33



What is hidden Markov model?

- **Markov chain**
  If fair and loaded coins are mixed!

0.5
0.5
H
T
0.5
0.5
0.6
0.4
H
T
0.6
0.4

$P_{F0} = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ [\,0.5 \;\; 0.5\,] \end{array}$

$P_F = \begin{array}{c|cc} & H & T \\ \hline H & 0.5 & 0.5 \\ T & 0.5 & 0.5 \end{array}$

Now we have to consider these probabilities, too!

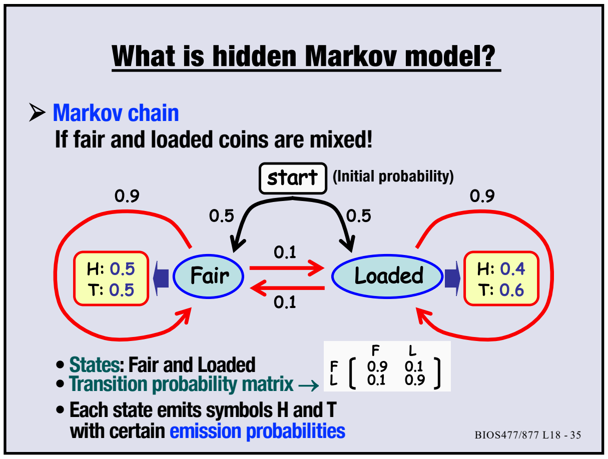
$P_{L0} = \begin{array}{c} \begin{array}{cc} H & T \end{array} \\ [\,0.4 \;\; 0.6\,] \end{array}$

$P_L = \begin{array}{c|cc} & H & T \\ \hline H & 0.4 & 0.6 \\ T & 0.4 & 0.6 \end{array}$

Observation: H, H, H

$P(H_FH_FH_F)$ or $P(H_LH_LH_L)$ or $P(H_FH_FH_L)$ or $P(H_FH_LH_L)$ or … ?

34



What is hidden Markov model?

- **Markov chain**
  If fair and loaded coins are mixed!

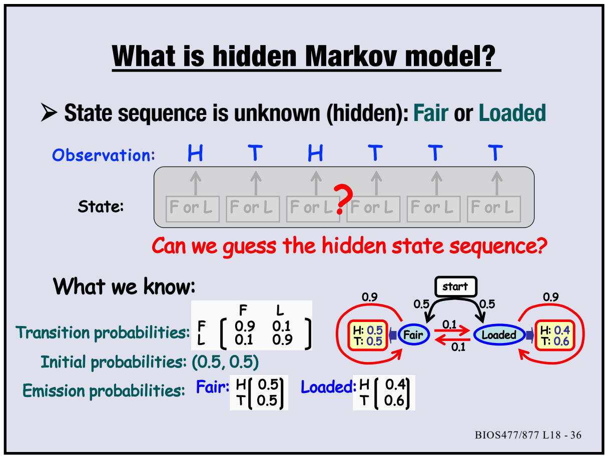
start
(Initial probability)
0.9
0.5
0.5
0.9
H: 0.5
T: 0.5
Fair
0.1
0.1
Loaded
H: 0.4
T: 0.6

- States: Fair and Loaded
- Transition probability matrix →

$\begin{array}{c|cc} & F & L \\ \hline F & 0.9 & 0.1 \\ L & 0.1 & 0.9 \end{array}$

- Each state emits symbols H and T with certain emission probabilities

35



What is hidden Markov model?

- State sequence is unknown (hidden): Fair or Loaded

Observation: H T H T T T
State:
F or L
F or L
F or L
F or L
F or L
F or L
?

Can we guess the hidden state sequence?

What we know:

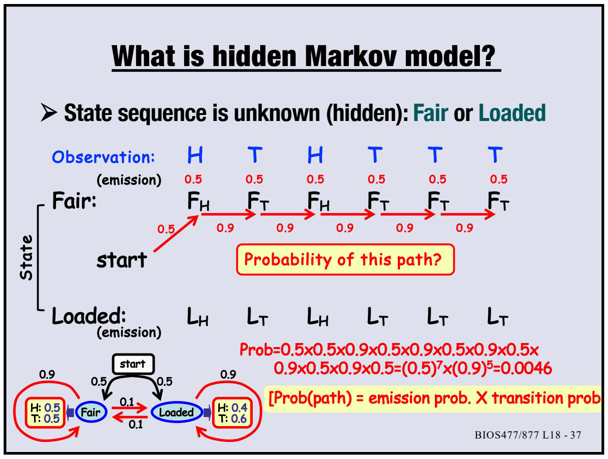
Transition probabilities: $\begin{array}{c|cc} & F & L \\ \hline F & 0.9 & 0.1 \\ L & 0.1 & 0.9 \end{array}$

Initial probabilities: (0.5, 0.5)

Emission probabilities: Fair: $\begin{array}{c|c} H & 0.5 \\ T & 0.5 \end{array}$ Loaded: $\begin{array}{c|c} H & 0.4 \\ T & 0.6 \end{array}$

start
0.9
0.5
0.5
0.9
H: 0.5
T: 0.5
Fair
0.1
0.1
Loaded
H: 0.4
T: 0.6

36



What is hidden Markov model?

- State sequence is unknown (hidden): Fair or Loaded

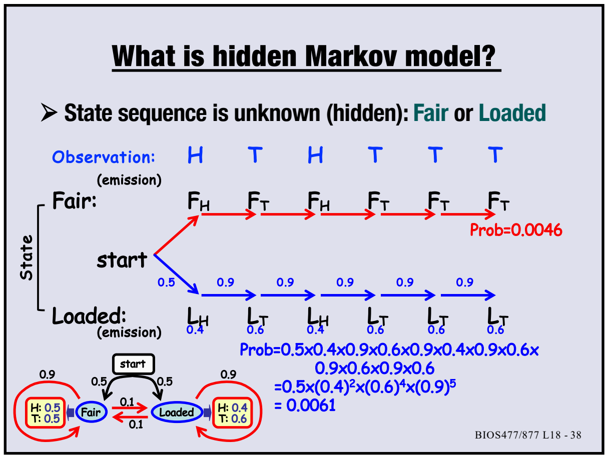
Observation: H T H T T T
(emission) 0.5 0.5 0.5 0.5 0.5 0.5
Fair: F_H F_T F_H F_T F_T F_T
0.5
0.9 0.9 0.9 0.9 0.9
State
start
Loaded: L_H L_T L_H L_T L_T L_T
(emission)
start
0.9
0.5
0.5
0.9
H: 0.5
T: 0.5
Fair
0.1
0.1
Loaded
H: 0.4
T: 0.6

Probability of this path?

Prob=0.5x0.5x0.9x0.5x0.9x0.5x0.9x0.5x
0.9x0.5x0.9x0.5=$(0.5)^7$x$(0.9)^5$=0.0046

[Prob(path) = emission prob. X transition prob

37



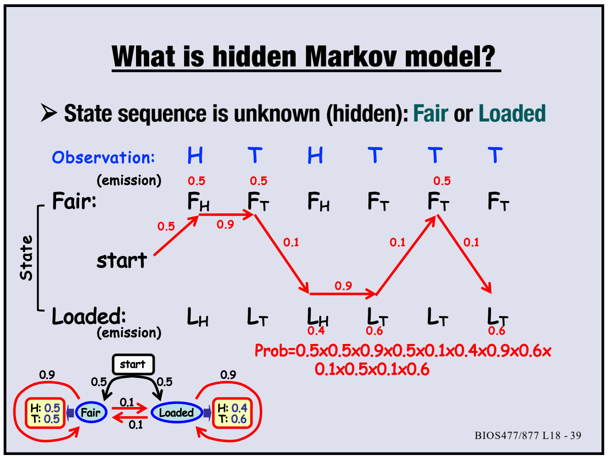
What is hidden Markov model?

- State sequence is unknown (hidden): Fair or Loaded

Observation: H T H T T T
(emission)
Fair: F_H F_T F_H F_T F_T F_T
Prob=0.0046
State
start
0.5
0.9 0.9 0.9 0.9 0.9
Loaded: L_H L_T L_H L_T L_T L_T
0.4 0.6 0.4 0.6 0.6 0.6
(emission)
start
0.9
0.5
0.5
0.9
H: 0.5
T: 0.5
Fair
0.1
0.1
Loaded
H: 0.4
T: 0.6

Prob=0.5x0.4x0.9x0.6x0.9x0.4x0.9x0.6x
0.9x0.6x0.9x0.6
=0.5x$(0.4)^2$x$(0.6)^4$x$(0.9)^5$
= 0.0061

38



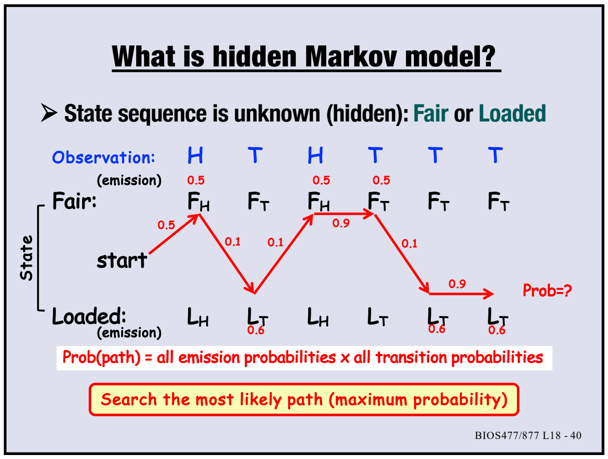
What is hidden Markov model?

➢ State sequence is unknown (hidden): Fair or Loaded

Prob=0.5x0.5x0.9x0.5x0.1x0.4x0.9x0.6x 0.1x0.5x0.1x0.6

BIOS477/877 L18 - 39

39



What is hidden Markov model?

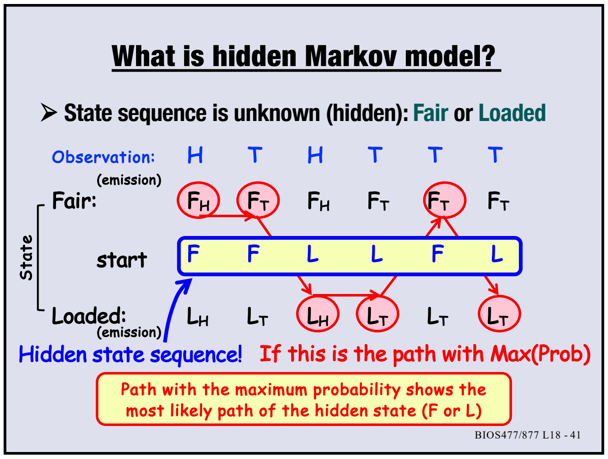
➢ State sequence is unknown (hidden): Fair or Loaded

Prob(path) = all emission probabilities x all transition probabilities

Search the most likely path (maximum probability)

BIOS477/877 L18 - 40

40



What is hidden Markov model?

➢ State sequence is unknown (hidden): Fair or Loaded

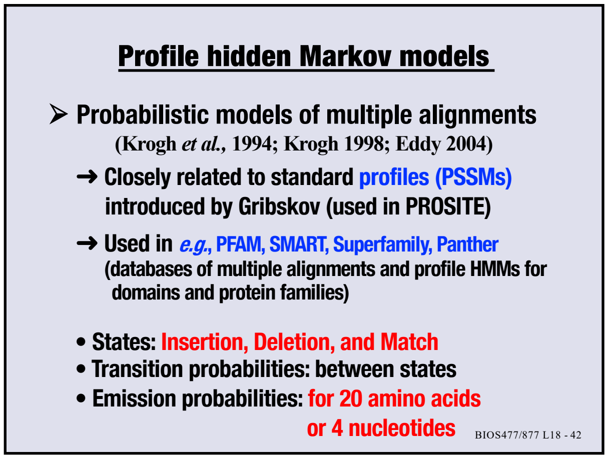
Observation: H T H T T T
(emission)
Fair: F_H F_T F_H F_T F_T F_T
State
start
F F L L F L
Loaded: L_H L_T L_H L_T L_T L_T
(emission)

Hidden state sequence! If this is the path with Max(Prob)

Path with the maximum probability shows the most likely path of the hidden state (F or L)

BIOS477/877 L18 - 41

41



Profile hidden Markov models

➢ Probabilistic models of multiple alignments (Krogh *et al.*, 1994; Krogh 1998; Eddy 2004)

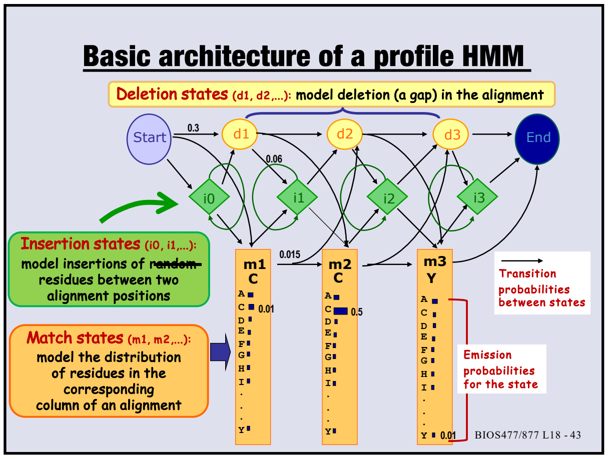
➔ Closely related to standard profiles (PSSMs) introduced by Gribskov (used in PROSITE)

➔ Used in *e.g.*, PFAM, SMART, Superfamily, Panther (databases of multiple alignments and profile HMMs for domains and protein families)

- States: Insertion, Deletion, and Match
- Transition probabilities: between states
- Emission probabilities: for 20 amino acids or 4 nucleotides

BIOS477/877 L18 - 42

42



Basic architecture of a profile HMM

Deletion states (d1, d2,...): model deletion (a gap) in the alignment

Insertion states (i0, i1,...): model insertions of ~~random~~ residues between two alignment positions

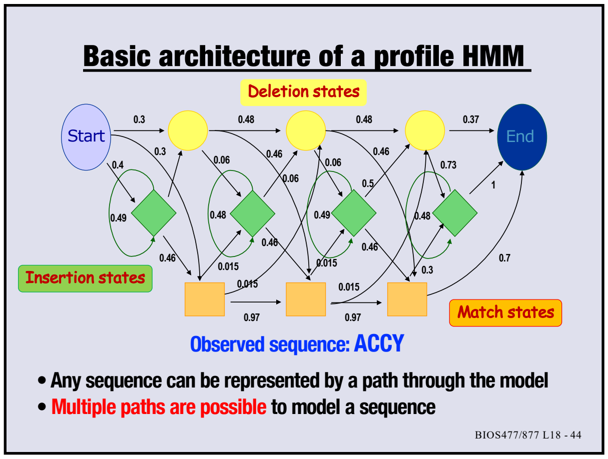
Match states (m1, m2,...): model the distribution of residues in the corresponding column of an alignment

Transition probabilities between states

Emission probabilities for the state

BIOS477/877 L18 - 43

43



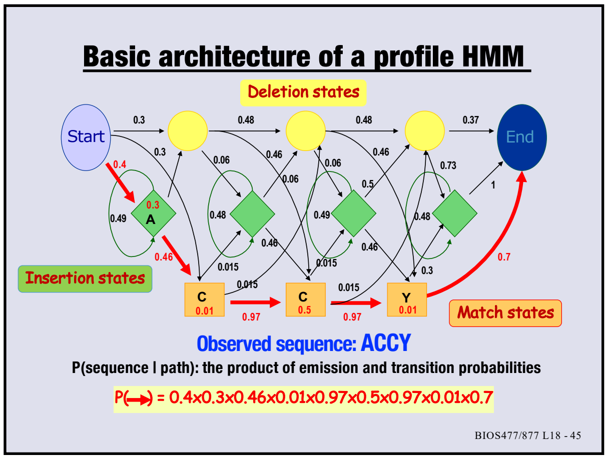
Basic architecture of a profile HMM

Observed sequence: ACCY

- Any sequence can be represented by a path through the model
- Multiple paths are possible to model a sequence

BIOS477/877 L18 - 44

44



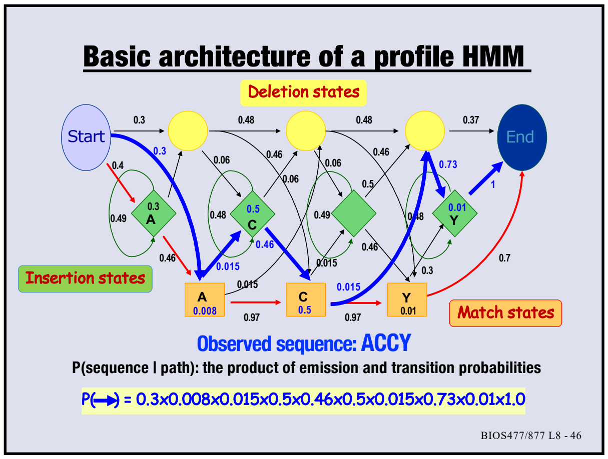
Basic architecture of a profile HMM

Observed sequence: ACCY

P(sequence I path): the product of emission and transition probabilities

P(→) = 0.4x0.3x0.46x0.01x0.97x0.5x0.97x0.01x0.7

45



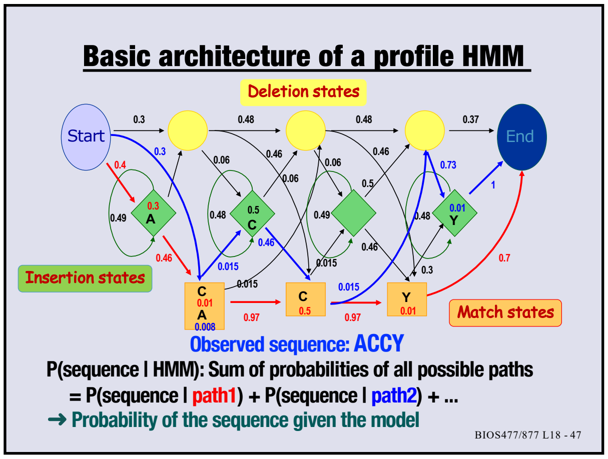
Basic architecture of a profile HMM

Observed sequence: ACCY

P(sequence I path): the product of emission and transition probabilities

P(→) = 0.3x0.008x0.015x0.5x0.46x0.5x0.015x0.73x0.01x1.0

46



Basic architecture of a profile HMM

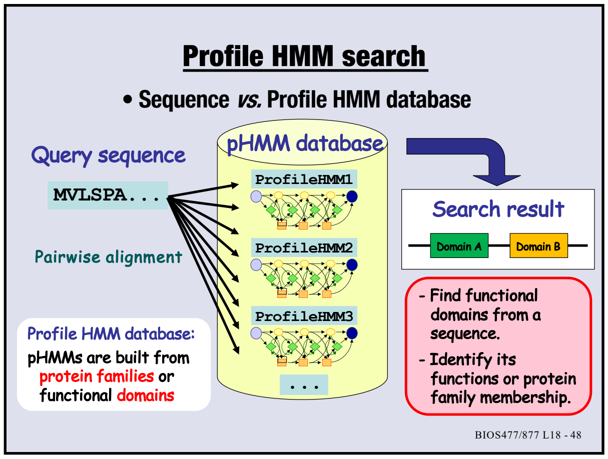
Observed sequence: ACCY

P(sequence I HMM): Sum of probabilities of all possible paths

= P(sequence I path1) + P(sequence I path2) + ...

➜ Probability of the sequence given the model

47



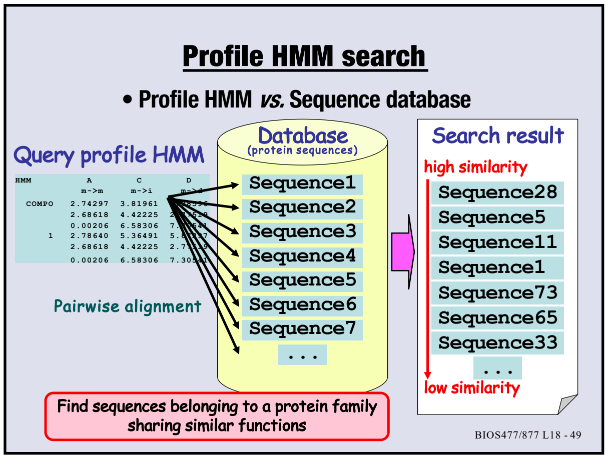
Profile HMM search

- Sequence *vs.* Profile HMM database

Profile HMM database: pHMMs are built from protein families or functional domains

- Find functional domains from a sequence.
- Identify its functions or protein family membership.

48



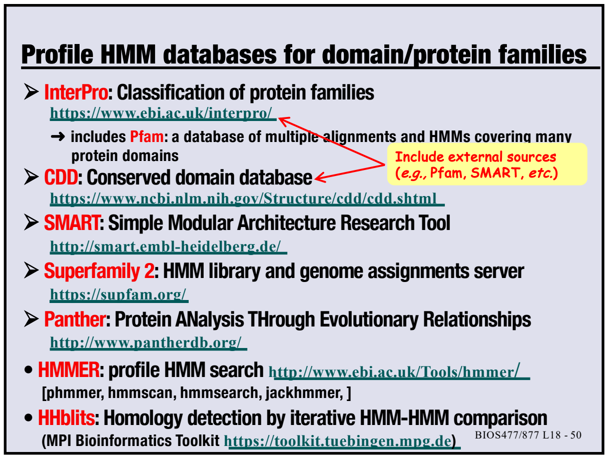
Profile HMM search

- Profile HMM *vs.* Sequence database

```
HMM          A        C        D
           m->m     m->i     m->d
  COMPO   2.74297  3.81961  [illegible]
          2.68618  4.42225  2.77519
          0.00206  6.58306  7.30541
      1   2.78640  5.36491  5.87297
          2.68618  4.42225  2.77519
          0.00206  6.58306  7.30541
```

Find sequences belonging to a protein family sharing similar functions

49



Profile HMM databases for domain/protein families

- ➢ InterPro: Classification of protein families
  https://www.ebi.ac.uk/interpro/
  ➜ includes Pfam: a database of multiple alignments and HMMs covering many protein domains
- ➢ CDD: Conserved domain database
  https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
- ➢ SMART: Simple Modular Architecture Research Tool
  http://smart.embl-heidelberg.de/
- ➢ Superfamily 2: HMM library and genome assignments server
  https://supfam.org/
- ➢ Panther: Protein ANalysis THrough Evolutionary Relationships
  http://www.pantherdb.org/

Include external sources (*e.g.*, Pfam, SMART, *etc.*)

- HMMER: profile HMM search http://www.ebi.ac.uk/Tools/hmmer/
  [phmmer, hmmscan, hmmsearch, jackhmmer,]
- HHblits: Homology detection by iterative HMM-HMM comparison
  (MPI Bioinformatics Toolkit https://toolkit.tuebingen.mpg.de)

50