Alignment statistics II / Algorithms II

Biol4230 Tues, February 13, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

Goals of today's lecture:

- what is the probability of an alignment score?
  - given two sequences
  - after a database search
  - after many database searches
- Hidden Markov Models
  - transition state models
  - profile HMMs

Inferring Homology from Statistical Significance

- Real *UNRELATED* sequences have similarity scores that are indistinguishable from *RANDOM* sequences
- If a similarity is NOT *RANDOM*, then it must be NOT *UNRELATED*
- Therefore, NOT *RANDOM* (statistically significant) similarity must reflect *RELATED* sequences

How often do things happen by chance? statistics of coin tosses - expectation

- p(H) = p(T) = 0.5
- p(HHHTH)=p(HTTTH)=p(HHHHH)=$(1/2)^5$
- how many times do we expect a run of 10 heads (by chance) in:

	Expectation	poisson probability
– 10 flips	1 $(1/2)^{10}$ = 0.001	0.001
– 100 flips	91$(1/2)^{10}$ ~ 0.1	0.1
– 1000 flips	991 $(1/2)^{10}$ ~ 1	0.6
– 1,000,000 flips	999,991 $(1/2)^{10}$ ~ 1000	0.999

- Probability (0 <= p <= 1) vs

  Expectation (0 <= E() <= number of trials)

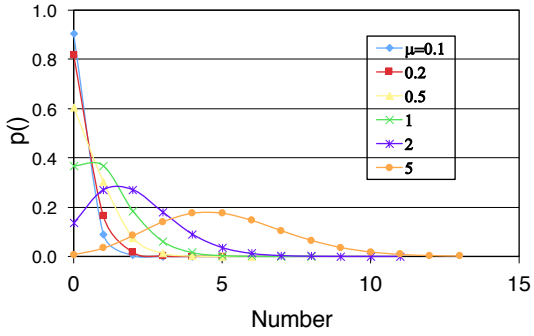
  E(x) = p(x) * N

Given an expectation, what is its probability?
The Poisson Distribution:
probabilities of counts of random events (radioactive decay, high similarity scores)

$$p(\mu,i) = \mu^i e^{-\mu} / i!$$

μ=mean expectation of event

i = number of events



Distribution of solitaire wins

- I play iphone solitaire compulsively when waiting
- I win about 25% of games
- If I have played 2,000 games, how many have I won? how often have I won 2 in a row, 3 in a row, etc.

in a row	p()	E(2000)
1	0.2	400
2	0.025	50
3	0.002	4
4	1e-4	0.3
5	6e-6	0.01

Poisson distribution for ranges of events (one or more)

$$p(x \geq 1) = \sum_{i=1}^{\infty} \mu^i e^{-\mu} / i! \;\; = \mu^1 e^{-\mu} /1! + \mu^2 e^{-\mu} /2! + \ldots$$

$$p(x \geq 1) = 1 - p(0) = 1 - \mu^0 e^{-\mu} /0!$$

$$= 1 - e^{-\mu}$$

μ	$p(x>0)$
0.001	0.001
0.01	0.010
0.1	0.095
1.0	0.632
2.0	0.865

$1\text{-}\exp^{-\mu} \sim \mu$ for $\mu < 0.1$

Statistics of "Head" runs

$$E(l) = n\,p^{\,l}$$

Results from tossing a coins 14 times; black circles indicate heads. The probability of 5 heads in a row is $p(5) = (1/2)^5 = 1/32$, but since there were 10 places that one could have obtained 5 heads in a row, the expected number of times that 5 heads occurs by chance is $E(5H) = 10 \times 1/32 = 0.31$.

Alignment scores as coin tosses

- E(# of H of length m) ~ np^m
- if the longest run is unique, $1 = np^{Rn}$

$$1/n = p^{Rn}$$

$$-\log_e(n) = R_n \log_e(p)$$

$$-\log_e(n)/\log_e(p) = R_n$$

$$R_n = \log_{(1/p)}(n)$$

Converting logarithms:

$10^x = B^y$

$x \log_{10} 10 = y \log_{10} B$

$x = y \log_{10} B$

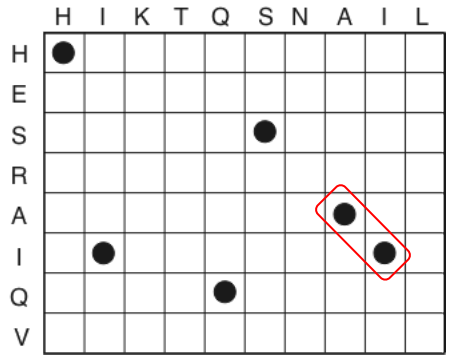
$x / \log_{10} B = y$

The expected length of the longest run R_n increases as log(n) of the run length

Statistics of "Head" alignments

$E(l) = m\ n\ p^{l}$

The expected length of the longest run R_n increases as log(mn).



Comparison of two protein sequences, with identities indicated as black circles. Assuming the residues were drawn from a population of 20, each with the same probability, the probability of an identical match is p = 0.05. In this example,there are m = 10 x n = 8 boxes, so E() = m n p = 80 x 0.05= 4 matches are expected by chance. The probability of two successive matches is $p^2 = (1/20)^2$ so a run of two matches is expected about n m $p^2 = 8 \times 10 \times (1/20)^2 = 0.2$ times by chance.

From "Head" runs to scores

The longest "Head" run is equivalent to the "longest hydrophobic stretch" using a scoring matrix that assigns positive values s_i for some residues *i* and –∞ for all other residues. Then:

$$p(S) = \sum p(s_i) \text{ for residues } i \text{ with } s_i > 0$$

The same analogy can be made for alignment scores between *i,j* where $s_{i,j}$ the score for aligning residues *i,j* is either + with $p(s_{i,j})$ or –∞ . Now the score for the longest positive alignment score is:

$$E(S \geq x) \propto mnp^{x}$$

$$E(S \geq x) \propto mne^{x \ln p}$$

$$E(S \geq x) \propto mne^{-\lambda x} \text{ where } \lambda = -\ln p$$

Karlin-Altschul statistics for alignments without gaps

Given:

$E(s_{i,j}) = \sum_{i,j} p_i p_j s_{i,j} < 0$ (local alignments)

Then:

$E(S \geq x) = Kmne^{-\lambda x}$

$K < 1$ (space correction)

λ solution of: $\sum_{i,j} p_i p_j e^{\lambda s_{i,j}}$

$E(S \geq x)$ is the Expectation (average # of times) of seeing score *S* in an alignment. so, we apply the Poisson conversion:

$p(x) = 1 - \exp(-x) \Rightarrow$

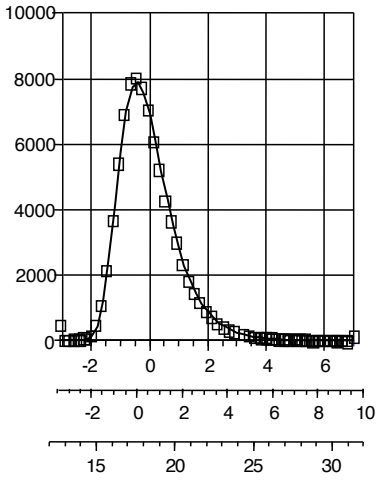
$p(S > x) = 1 - \exp(-Kmne^{-\lambda S})$

The Similarity Statistics Mantra…

- Find the *Probability* of a rare event (e.g. a high score) in a cluster of residues $p^n \propto e^{-\lambda S}$
- Find the *Expectation* of this event by correcting for all the places it could have happened $Kmn \bullet e^{-\lambda S}$
- Convert that into a *Probability* using the Poisson formula: $1 - \exp(-Kmne^{-\lambda S})$
- Convert that *Probability* into an *Expectation* for the number of sequences in the database

$$E(S > x) = P \bullet D = (1 - \exp(-Kmne^{-\lambda S})) \bullet D$$

Extreme value distribution



$S' = \lambda S_{raw} - \ln K\, m\, n$

$S_{bit} = (\lambda S_{raw} - \ln K)/\ln(2)$

$P(S' > x) = 1 - \exp(-e^{-x})$

$P(S_{bit} > x) = 1 - \exp(-mn2^{-x})$

$E(S' > x \mid D) = P\, D$

$P(B \text{ bits}) = m\, n\, 2^{-B}$

$P(40 \text{ bits}) = 1.5 \times 10^{-7}$

$E(40 \mid D=4000) = 6 \times 10^{-4}$

$E(40 \mid D=50E6) = 7.5$

How many bits do I need?

$$P(S_b > x_b) = mn2^{-x_b} = \frac{mn}{2^{x_b}}, \; S_b \text{ is a score in "bits"}$$

Query size m	Lib. seq. size: n	DB Entries D	mnD/0.01	Bit threshold
200	200	100,000	$4 \times 10^9/0.001$	42
450	450	100,000	$2 \times 10^{10}/0.001$	44
450	450	10,000,000	$2 \times 10^{13}/0.001$	51

How many “bits” do I need?

E(p I D) = p(40 bits) x database size

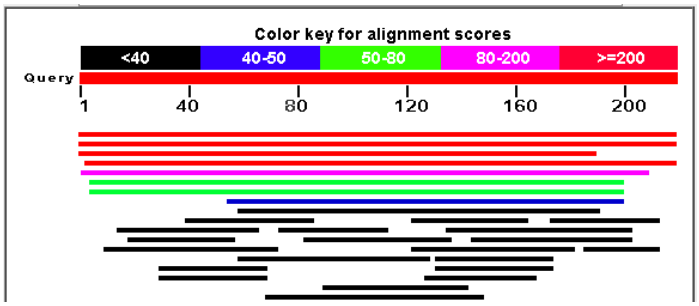
$E(40 \mid 4{,}000) = 10^{-8} \times 4{,}000 = 4 \times 10^{-5}$	(significant)
$E(40 \mid 40{,}000) = 10^{-8} \times 4 \times 10^{4} = 4 \times 10^{-4}$	(significant)
$E(40 \mid 400{,}000) = 10^{-8} \times 4 \times 10^{5} = 4 \times 10^{-3}$	(not significant)

To get E() ~ 10^{-3} :

genome (10,000) p ~ $10^{-3}/10^{4} = 10^{-7}/160{,}000 = 40$ bits

SwissProt (500,000) p ~ $10^{-3}/10^{6} = 10^{-9}/160{,}000 = 47$ bits

Uniprot/NR (10^{7}) p ~ $10^{-3}/10^{7} = 10^{-10}/160{,}000 =$ 50 bits



very significant 10^{-50}

significant 10^{-6}

significant 10^{-3}

not significant

Statistics, validation, HMMs

- what is the probability of an alignment score?
  - given two sequences
  - after a database search
  - after many database searches
- Hidden Markov Models
  - transition state models
  - profile HMMs
  - HMMER3

Should you trust the E()-value??

- The inference of homology from statistically significant similarity depends on the observation that unrelated sequences look like random sequences
  - Is this ALWAYS true?
  - How can we recognize when it is not true?
- If unrelated==random, then the E()-value of the highest scoring unrelated sequence should be E() ~ 1.0
- Statistical estimates can also be confirmed by searches against shuffled sequences

Smith-Waterman (ssearch36)
– highest scoring unrelated from domains

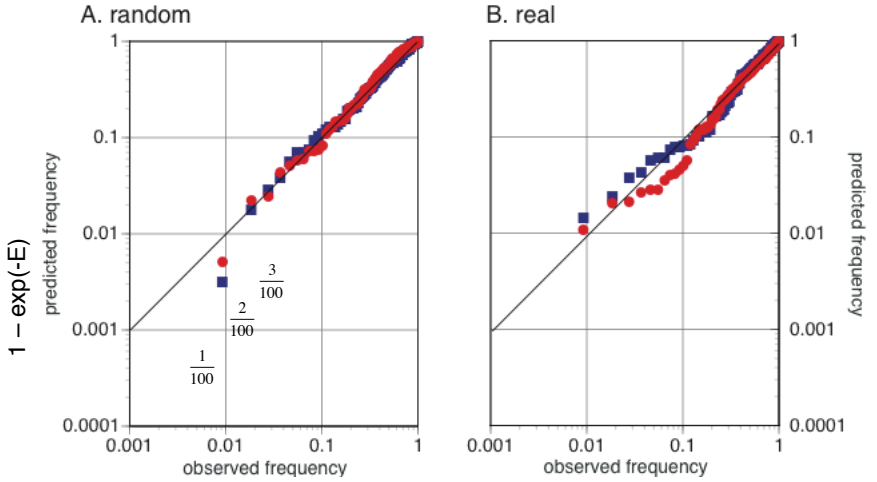
sp|P46420.2|GSTF4_MAIZE	Glutathione S-transferase 4; GS (223)	74	34.2	1.1	0.236	0.500	212	align		
sp|Q13155.2|AIMP2_HUMAN	Aminoacyl tRNA synthase complex (320)	73	33.7	2.3	0.349	0.674	43	align		
sp|P46421.1|GSTU5_ARATH	Glutathione S-transferase U5; (224)	71	33.0	2.6	0.217	0.580	143	align		
sp|Q9SR36.1|GSTU8_ARATH	Glutathione S-transferase U8; (224)	71	33.0	2.6	0.279	0.596	104	align		
sp|P13860.1|GUX1_PHACH	Exoglucanase 1; 1,4-beta-cellobi (516)	74	34.0	3	0.327	0.615	52	align		
sp|P0A9D3.1|GSTA_ECO57	Glutathione S-transferase GstA g (201)	70	32.6	3	0.276	0.529	87	align		
sp|P26641.3|EF1G_HUMAN	Elongation factor 1-gamma; EF-1 (437)	73	33.7	3.2	0.268	0.575	127	align		
sp|Q9LZI9.1|GSTFD_ARATH	Glutathione S-transferase F13; (219)	70	32.6	3.3	0.265	0.547	117	align		
sp|Q2NL00.3|GSTT1_BOVIN	Glutathione S-transferase theta (240)	70	32.6	3.7	0.362	0.638	47	align		
sp|Q29387.2|EF1G_PIG	Elongation factor 1-gamma; EF-1-g (432)	72	33.3	4.2	0.268	0.567	127	align		
sp|Q759Q6.1|DAD2_ASHGO	DASH complex subunit DAD2; Outer (111)	66	31.2	4.6	0.312	0.667	48	align		
sp|Q61133.4|GSTT2_MOUSE	Glutathione S-transferase theta (244)	69	32.2	5	0.193	0.545	176	align		
sp|P78417.2|GSTO1_HUMAN	Glutathione S-transferase omega (241)	68	31.8	6.4	0.271	0.588	85	align		

The highest scoring unrelated sequence should have an E()-value ~ 1
In one search.

What about after 10 searches?
After 100?
After 10,000?

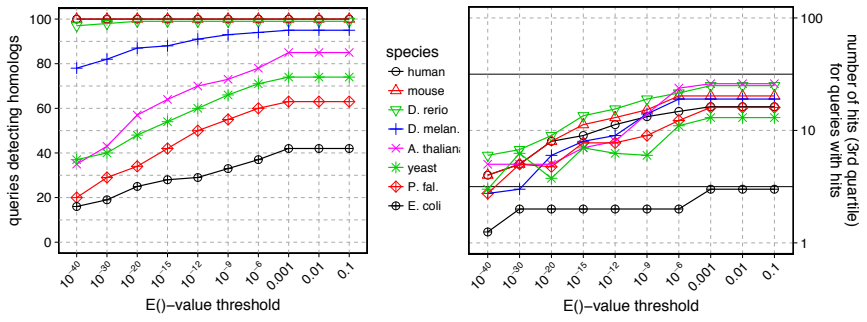
Expectations are turned into probabilities using: 1 – exp(-E)

Highest unrelated E()values decrease with more searches



correct for multiple searches

Detectable homologs to human enzymes varying E()-value threshold



E()-values when??

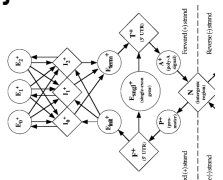
- E()-values (BLAST expect) provide accurate statistical estimates of similarity by chance
  - non-random -> not unrelated (homologous)
  - E()-values are accurate (0.001 happens 1/1000 by chance)
  - E()-values factor in (and depend on) sequence lengths and database size
- E()-values are NOT a good proxy for evolutionary distance
  - doubling the length/score SQUARES the E()-value
  - percent identity (corrected) reflects distance (given homology)

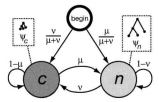
Statistics, validation, HMMs

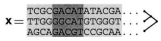
- what is the probability of an alignment score?
  - given two sequences
  - after a database search
  - after many database searches
- Hidden Markov Models
  - transition state models
  - profile HMMs
  - HMMER2

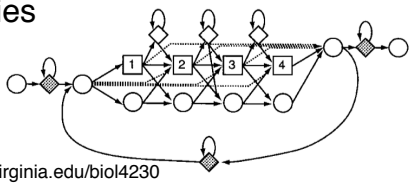
Why HMMs (Hidden Markov Models) ?

- HMMs provide a general purpose strategy for fitting models with adjacent features to data
  - gene models:
    genscan/twinscan
  - conserved regions:
    phastcons
  - protein domain families
    profile HMMs
    hmmer/pfam









profile-HMMs – Used by Pfam

- Anders Krogh in David Haussler's group.
- Takes the "standard" profiles and uses HMM based "standard" mathematics to solve two problems
  - Profile-HMM scores are comparable (*)
  - Setting gap costs
- Theoretical framework for what we are doing.
- (* this is not really true. see later)

A simple Hidden Markov Model

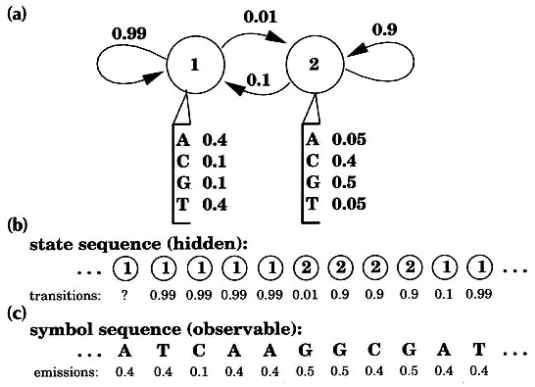


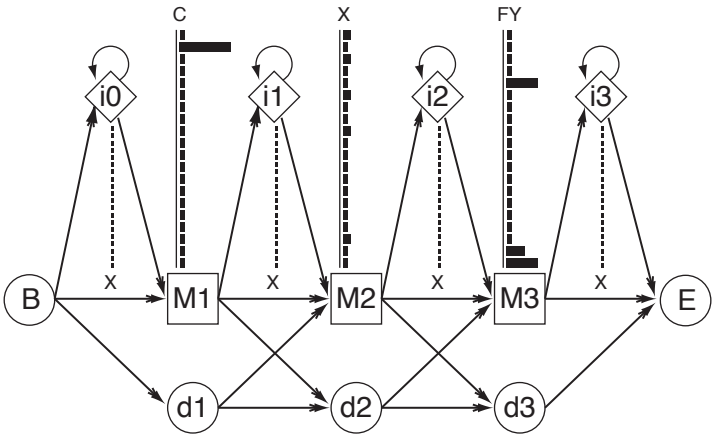
Figure 1 A simple hidden Markov model. A two-state HMM describing DNA sequence with a heterogeneous base composition is shown, following work by Churchill [10]. (a) State 1 (top left) generates AT-rich sequence, and state 2 (top right) generates CG-rich sequence. State transitions and their associated probabilities are indicated by arrows, and symbol emission probabilities for A,C,G and T for each state are indicated below the states. (For clarity, the begin and end states and associated state transitions necessary to model sequences of finite length have been omitted.) (b) This model generates a state sequence as a Markov chain and each state generates a symbol according to its own emission probability distribution (c). The probability of the sequence is the product of the state transitions and the symbol emissions. For a given observed DNA sequence, we are interested in inferring the hidden state sequence that 'generated' it, that is, whether this position is in a CG-rich segment or an AT-rich segment.

Eddy, S. R. Hidden Markov models. *Curr Opin Struct Biol* **6,** 361–365 (1996).

Profile (protein family) HMMs

1 2 3

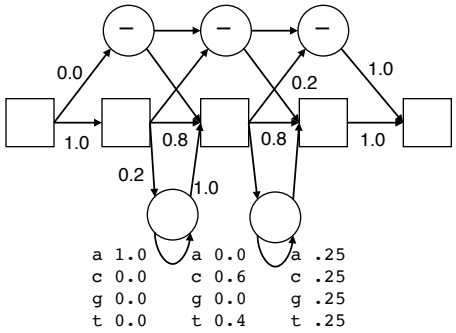
C A F
C G W
C D Y
C V F
C K Y



Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).

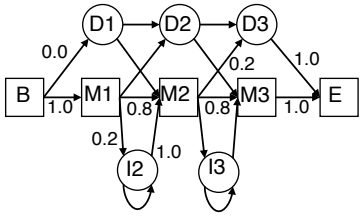
HMM transitions and emissions are probabilities

```
a - c g
a - t a
a - c c
a t t t
a - c -
```



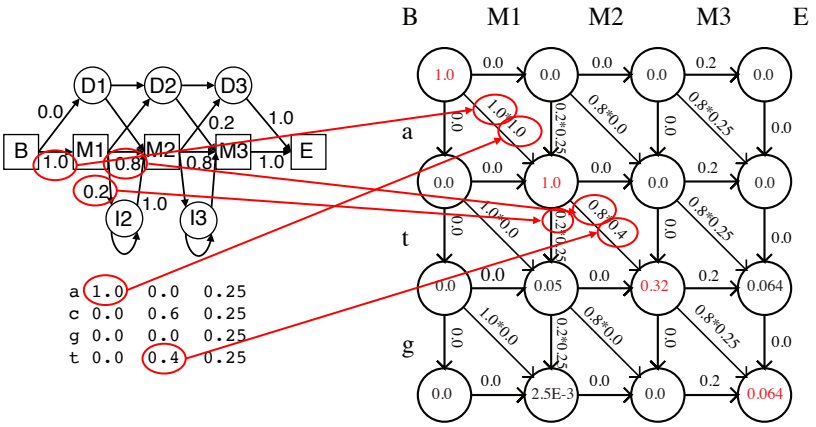
Given an HMM – how do we calculate a score (assuming an alignment)?

```
a - c g
a - t a
a - c c
a t t t
a - c -
```

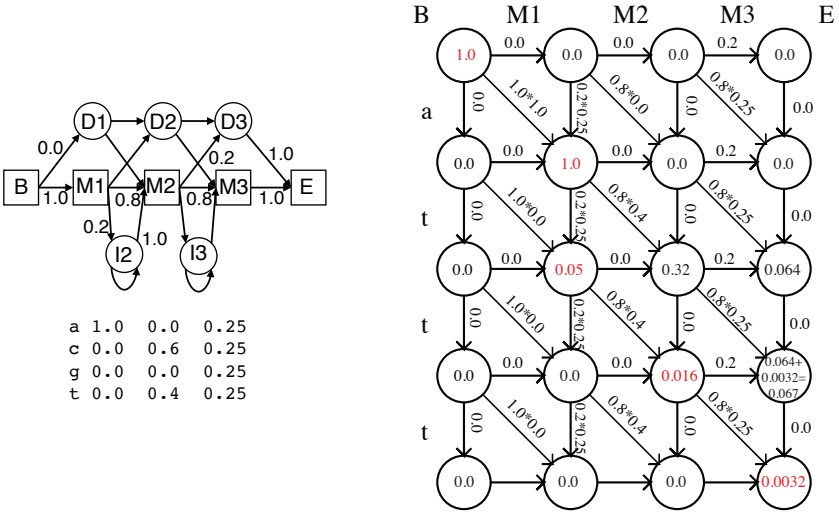


$$p(atg|HMM) = p(B)p(M1|B)p(a|M1)p(M2|M1)p(t|M2)p(M3|M2)p(g|M3)p(E|M3)$$
$$= 1.0 * 1.0 * 1.0 * 0.8 * 0.4 * 0.8 * 0.25 * 1.0 = 0.064$$
$$p(attt|HMM) =$$
$$p(B)p(M1|B)p(a|M1)p(I2|M1)p(t|I2)p(M2|I2)p(t|M2)p(M3|M2)p(g|M3)p(E|M3)$$
$$= 1.0 * 1.0 * 1.0 * 0.2 * 0.25 * 1.0 * 0.4 * 0.8 * 0.25 * 1.0 = 0.004$$

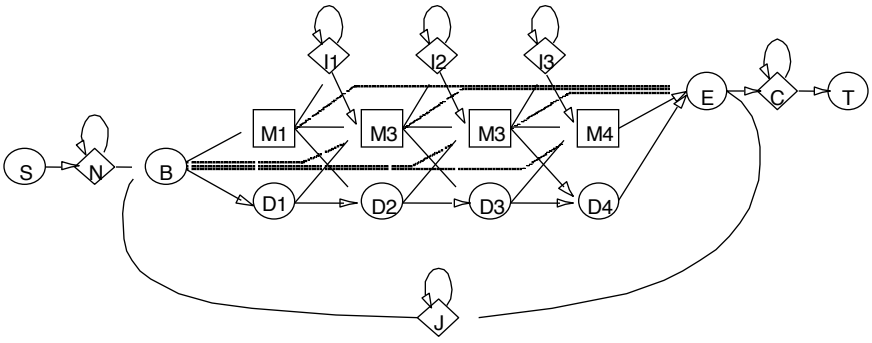
HMM – finding the best alignment dynamic programming



HMM – alignment with dynamic programming



HMMER- ‘Plan 7’ profile HMM



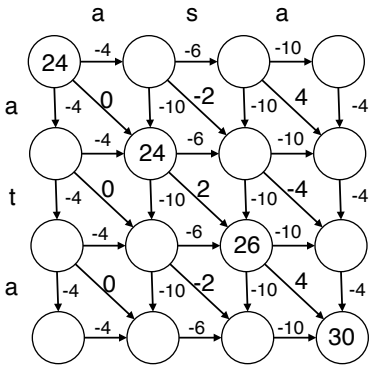
Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).

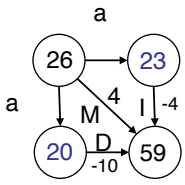
HMM Algorithms

1. The scoring problem: P(seq | model)
   "Forward" algorithm
   (sums over all alignments)
2. The alignment problem: max P(seq, statepath | model)
   "Viterbi" algorithm
3. The training problem:
   "Forward-backward" algorithm and
   Baum-Welch expectation maximization

For profile HMMs, all three algorithms use O(MN) dynamic programming -- same as "standard" Smith/Waterman and Needleman/Wunsch.

HMM Alignment





30+10+19

$$F_j^M(i) = \log \frac{e_{M_j}(x_i)}{q_{x_i}} + \log[a_{M_{j-1}M_j} \exp(F_{j-1}^M(i-1))$$
$$+ a_{I_{j-1}M_j} \exp(F_{j-1}^I(i-1)) + a_{D_{j-1}M_j} \exp(F_{j-1}^D(i-1))]$$

Needleman-Wunsch
max log likelihood
HMM Viterbi alignment

HMM Forward (score)

Σ probabilities

hmmbuild –
from multiple sequence alignment to hmm

```
CLUSTAL 2.0.12 multiple sequence alignment

GSTP1_HUMAN      --MPPYTVVYFPVRGRCAALRMLLADQGQSWKEEVVTV--------ETWQEGSLKASCL
GSTM1_HUMAN      ----MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLD
GSTM3_HUMAN      MSCESSMVLGYWDIRGLAHAIRLLLEFTDTSYEEKRYTCGEAPDYDRSQWLDVKFKLDLD
GSTA1_HUMAN      --MAEKPKLHYFNARGRMESTRWLLAAAGVEFEEKFIKS-------AEDLDKLRNDGYLM
                         : *:  **   : * **   . .::*:  .        .   .   .
...
GSTP1_HUMAN      PGCLDAFPLLSAYVGRLSARPKLKAFLASPEYVNLPINGNGKQ--------
GSTM1_HUMAN      PKCLDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK----
GSTM3_HUMAN      PKCLDEFPNLKAFMCRFEALEKIAAYLQSDQFCKMPINNKMAQWGNKPVC-
GSTA1_HUMAN      SSLISSFPLLKALKTRISNLPTVKKFLQPGSPRKPPMDEKSLEEARKIFRF
                 .  :. ** *.    *:.   .:  :: .      *:  :
```

```
HMM      A        C        D        E        F        G        H        I        W        Y
       m->m     m->i     m->d     i->m     i->i     d->m     d->d
COMPO 2.61963  4.31739  2.89583  2.62705  3.16314  3.03683  3.80746  2.80705  4.63822  3.29333
      2.68622  4.42229  2.77523  2.73127  3.46358  2.40517  3.72498  3.29358  4.58481  3.61507
      0.49776  2.03151  1.34335  0.66196  0.72534  0.00000        *
  1   2.61925  2.59613  4.05856  3.53413  3.26650  3.61183  4.19513  2.30607  4.93453  3.72168  3 l - - -
      2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  4.58477  3.61503
      0.03191  3.85649  4.57884  0.61958  0.77255  0.51074  0.91641
  2   2.06827  4.54009  3.12380  2.21293  3.75914  3.45042  3.76301  3.02955  5.15348  3.87801  4 a - - -
      2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  4.58477  3.61503
      0.02682  4.02764  4.74999  0.61958  0.77255  0.41306  1.08359
  3   2.61989  4.76650  2.97682  2.05462  4.02949  3.42092  3.68173  3.43295  5.31354  3.98992  5 e - - -
      2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  4.58477  3.61503
      0.02373  4.14859  4.87094  0.61958  0.77255  0.48576  0.95510
```

20 amino acids
7 transitions

-ln(p)

HMMR3.1 – jackhmmer: psiblast with HMMs

```
# jackhmmer :: iteratively search a protein sequence against a protein database
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query sequence file:             mgstm1.aa
# target sequence database:        /slib2/fa_dbs/pir1.lseg
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Query:       sp|P10649|GSTM1_MOUSE  [L=218]
Description: Glutathione S-transferase Mu 1; GST 1-1; GST class-mu 1;
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence
    -------  ------ -----    -------  ------ -----   ---- --  --------
+  1.4e-124  413.3   1.7   1.6e-124  413.2   1.7    1.0  1  sp|P08010|GSTM2_RAT
+   8.3e-25   87.1   0.0    1.2e-24   86.6   0.0    1.1  1  sp|P09211|GSTP1_HUMAN
+     4e-23   81.6   0.0    5.6e-23   81.1   0.0    1.1  1  sp|P04906|GSTP1_RAT
+   1.6e-14   53.5   0.3      2e-14   53.2   0.3    1.1  1  sp|P00502|GSTA1_RAT
+     1e-08   34.5   0.1    1.5e-08   34.0   0.1    1.2  1  sp|P14942|GSTA4_RAT
+   0.00028   20.0   0.0       0.15   11.1   0.0    2.5  3  sp|P04907|GSTF3_MAIZE
------ inclusion threshold ------
     0.0031   16.6   0.0     0.0061   15.6   0.0    1.5  1  sp|P12653|GSTF1_MAIZE
```

http://hmmr.org/

HMMR3.1 – jackhmmer: iteration 2

```
@@
@@ Round:                  2
@@ Included in MSA:        7 subsequences (query + 6 subseqs from 6 targets)
@@ Model size:             218 positions
@@
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence
    -------  ------ -----    -------  ------ -----   ---- --  --------
   1.5e-111  370.7   0.2   1.7e-111  370.5   0.2    1.0  1  sp|P08010|GSTM2_RAT
    8.5e-92  306.1   0.0    1.1e-91  305.7   0.0    1.0  1  sp|P04906|GSTP1_RAT
    3.1e-90  301.0   0.0    4.2e-90  300.6   0.0    1.0  1  sp|P09211|GSTP1_HUMAN
    3.1e-84  281.4   0.5    3.6e-84  281.2   0.5    1.0  1  sp|P00502|GSTA1_RAT
    2.2e-74  249.2   0.0    2.8e-74  248.8   0.0    1.0  1  sp|P14942|GSTA4_RAT
    1.9e-17   63.0   0.0    2.3e-11   43.2   0.0    2.0  2  sp|P04907|GSTF3_MAIZE
+   2.7e-17   62.6   0.0    3.5e-17   62.2   0.0    1.2  1  sp|P12653|GSTF1_MAIZE
+   3.6e-08   32.7   0.0    4.5e-08   32.4   0.0    1.1  1  sp|P20432|GSTT1_DROME
+   0.00016   20.8   0.0     0.0011   18.0   0.0    2.0  1  sp|P0ACA5|SSPA_ECO57
------ inclusion threshold ------
      0.078   12.0   0.1         11    5.0   0.0    3.4  2  sp|P07814|SYEP_HUMAN
```

http://hmmr.org/

HMMER3.1 alignments w/ confidence limits

```
>> sp|P20432|GSTT1_DROME  Glutathione S-transferase 1-1; DDT-dehydrochlorinase; GST class-theta
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !   32.4   0.0   3.4e-11   4.5e-08      54     169 ..      47     169 ..       2     183 .. 0.72

  Alignments for each domain:
  == domain 1  score: 32.4 bits;  conditional E-value: 3.4e-11
                  xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx....xxxxxxxxxxxxxxxxxxxxxxx.xxxxxxxxxxx..xx RF
GSTM1_MOUSE-i1  54 gllfgqlPlliDGdlkltqsrailrylarkyn....lyGkdekerirvDmvedgveDlrlk.lislvykpdfek..ek 124
                      +P+l+D    l +srai  yl +ky+    ly k  k r+ ++      + +  + +++  y+  f k  ++
sp|GSTT1_DROME  47 INPQHTIPTLVDNGFALWESRAIQVYLVEKYGktdsLYPKCPKKRAVINQRLYFDMGTLYQsFANYYYPQVFAkapAD 124
                  3355689*****99**************99964444899999999998654444444444045555655656652246 PP

                  xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx RF
GSTM1_MOUSE-i1 125 deylkalpeklklfeklLgkkaflvGnkisyvDillldlllvvev 169
                  +e+ k++++  +  +++L+++++ +G+ ++ +Di l+  + ++ev
sp|GSTT1_DROME 125 PEAFKKIEAAFEFLNTFLEGQDYAAGDSLTVADIALVATVSTFEV 169
                  88999999999999**********************999888876 PP
```


HMMER3.1 – domain output

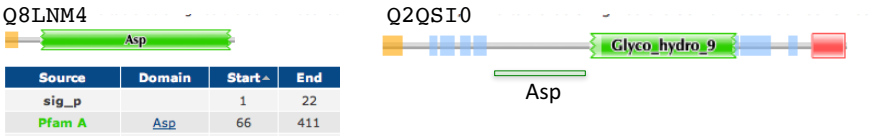
```
>> sp|P04907|GSTF3_MAIZE  Glutathione S-transferase 3; GST class-phi member 3; GST-III
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !   43.2   0.0   1.8e-14   2.3e-11      40      91 ..      35      86 ..      16      93 .. 0.86
   2 !   17.9   0.0   9.2e-07    0.0012     127     196 ..     136     207 ..     126     214 .. 0.87

  Alignments for each domain:
  == domain 1  score: 43.2 bits;  conditional E-value: 1.8e-14
                  xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx RF
GSTM1_MOUSE-i1 40 dldreqwlkeklklgllfgqlPlliDGdlkltqsrailrylarkynlyGkde 91
                  dl      + ++ +   fgq+P+l+DGd++l++srai+ry+a+ky+++G d
sp|GSTF3_MAIZE 35 DLTTGAHKQPDFLALNPFGQIPALVDGDEVLFESRAINRYIASKYASEGTDL 86
                  6666667778888889**********************************985 PP

 domain 2  score: 17.9 bits;  conditional E-value: 9.2e-07
                  xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx..xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx RF
GSTM1_MOUSE-i1 127 ylkalpeklklfeklLgkkaflvGnkisyvDil..lldlllvvevlepkvLdaFPlLkafvaRlsalpkikk 196
                    +++l + l ++e  L  +++l+G+  + +D +  ll +l   +   p+++ a P +ka+   + a+p  +k
sp|GSTF3_MAIZE 136 HAEQLAKVLDVYEAHLARNKYLAGDEFTLADANhaLLPALTSARPPRPGCVAARPHVKAWWEAIAARPAFQK 207
                  55677777999*****************99754499************************9999998776 PP
```

Improving sensitivity with protein/domain family models

- HMMER3 – jackhmmer – method
  1. do HMMER (Hidden Markov Model, HMM) search with single sequence
  2. use query-HMM-based implied multiple sequence alignment to more accurate HMM
  3. repeat steps 1 and 2 with HMM
- HMMER3– results:
  1. Less over-extension because of probabilistic alignment
  2. Used to construct Pfam domain database
     - Many protein families are too diverse for one HMM, Pfam divides families into multiple HMMs and groups in Clans
  3. Clearly homologous sequences are still missed

Missing homology beyond the HMM model

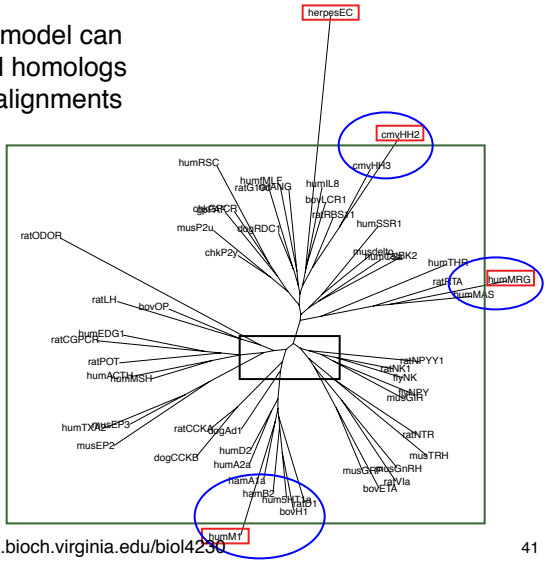
```
>>tr|Q8LNM4|Q8LNM4_ORYSJ Eukaryotic aspartyl protease family protein vs
>>tr|Q2QSI0|Q2QSI0_ORYSJ Glycosyl hydrolase family 9 protein, expressed OS=O  (694 aa)
 qRegion: 134-277:172-311 : score=508; bits=240.8; LPr=67.0 : Aspartyl protease
 s-w opt: 508  Z-score: 1248.7  bits: 240.8 E(1): 9.6e-68
Smith-Waterman score: 508; 62.5% identity (79.2% similar) in 144 aa overlap

           130       140       150       160       170       180       190       200
Q8LNM4 TDACKSIPTSNCSSNMCTYEGTINSKLGGHTLGIVATDTFAIGTATASLGFGCVVASGIDTMGGPSGLIGLGRAPSSLVS
                  ::: :.: :: . :.   : : : : :::::.::: :.: ::::::: :::: : ::..::::.: :::.
Q2QSI0 LCESISNDIHNCSGNVCMYEASTNA---GDTGGKVGTDTFAVGTAKANLAFGCVVASNIDTMDGSSGIVGLGRTPWSLVT
            170       180           190       200       210       220       230
           210       220       230       240       250       260       270       280
Q8LNM4 QMNITKFSYCLTPHDSGKNSRLLLGSSAKLAGGGNSTTTPFVKTSPGDDMSQYYPIQLDGIKAGDAAIALPPSGNTVLVQ
       :  .. :::::.:::.:::. :.::::.::::::: ...::::  : :.:.: :: .::. .:::::  : :::::
Q2QSI0 QTGVAAFSYCLAPHDAGKNNALFLGSTAKLAGGGKTASTPFVNIS-GNDLSNYYKVQLEVLKAGDAMIPLPPSGVLWDNY
      240       250       260       270       280       290       300       310
```



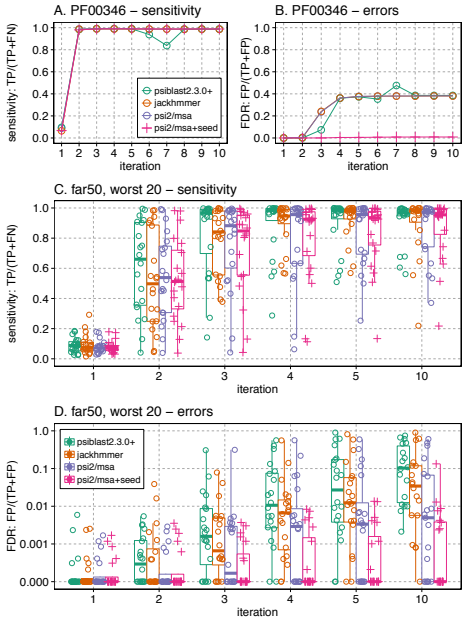
Source	Domain	Start	End
sig_p		1	22
Pfam A	Asp	66	411

Pfam misses/mis-aligns proteins distant from the model

- For diverse families, a single model can find, and miss, closely related homologs
- Even if homologs are found, alignments may be short



How much improvement with PSSMs/ HMMs?



Pearson (2017) Nuc. Acids Res. 45:e46

Statistics, validation, HMMs

- what is the probability of an alignment score?
  - given two sequences
    - probability of match, times number of match run starts: extreme value
  - after a database search
    - Bonferroni correction for database size
  - after many database searches
    - Bonferroni correction for number of searches (?)
    - what happens to false negatives?
- Hidden Markov Models
  - transition state models
  - profile HMMs
  - HMMER3
    - better, but sometimes missed
    - How might one find “missing” homologs?