Mathematical and Computational Molecular Biology
Math/Biol 227 – Brendel – WQ94/95

Time & Location: Tu, Th 1:15P - 2:30P (**Units:** 3); Gilbert 119
Instructor: Volker Brendel (380-383A; 723-9256; volker@grendel.stanford.edu)

Synopsis

Precipitated by an enormous increase in molecular sequence data (both DNA and protein), computational tools have become essential to molecular biology research. This course seeks to provide a state-of-the-art introduction to the subject. Emphasis will be on concepts and principles, combined with hands-on (-keyboard) applications.

Prerequisites

This interdisciplinary course is directed at graduate and interested undergraduate students of biology, medicine, computer science, statistics, operations research, or mathematics. There are no formal prerequisites as all necessary knowledge will be developed in the course. However, some knowledge of the fundamental concepts of molecular biology and statistical analysis will be helpful. See the instructor for any questions regarding this.

Mathematical and Computational Molecular Biology
Math/Biol 227 – Brendel – WQ94/95

Time & Location: Tu, Th 1:15P - 2:30P (**Units:** 3); Gilbert 119
Instructor: Volker Brendel (380-383A; 723-9256; volker@grendel.stanford.edu)

Tentative Agenda

I **Tu, Jan. 10** Overview.

II **Th, Jan. 12** Pairwise sequence comparisons I; algorithms of Needleman & Wunsch and Smith & Waterman.

III **Tu, Jan. 17** Pairwise sequence comparisons II; gap penalties; suboptimal alignments.

IV **Th, Jan. 19** Score-based sequence analysis; single sequence features.

V **Tu, Jan. 24** Score-based sequence analysis; pairwise sequence comparisons (SSPA).

VI **Th, Jan. 26** Amino acid substitution scoring matrices.

VII **Tu, Jan. 31** Query search methods; FASTA, BLAST.

VIII **Th, Feb. 2, S. Karlin instructor** Sequence comparisons not requiring alignments.

IX **Tu, Feb. 7** Phylogenetic trees from sequence data I.

X **Th, Feb. 9** Phylogenetic trees from sequence data II.

XI **Tu, Feb. 14** Runs, patterns, clusters of particular letter types.

XII **Th, Feb. 16, J. Kleffe instructor** Word counts.

XIII **Tu, Feb. 21** Spacings between sequence markers.

XIV **Th, Feb. 23** Profile methods.

XV **Tu, Feb. 28** Optimal signal search profiles (EM algorithm).

XVI **Th, March 2** Hidden Markov chain Models.

XVII **Tu, March 7, D. Brutlag instructor** Belief systems and neural networks for secondary structure prediction.

XVIII **Th, March 9** Gene and intron prediction.

XIX **Tu, March 14** Dead week lecture I.

XX **Th, March 16** Dead week lecture II.

1/10/95

Math & Computer Molecular Biology - Intro

I) Intro

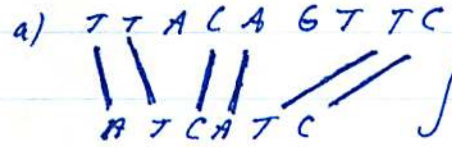
A) Sequences
- DNA
- protein

B) Databases
- Genbank, Genpept
- EMBL, Swiss-prot

Nature 349:99. Gilbert

II) Data-based
Package-based
Technique-based

III) Pairwise-Sequence Comparisons

a) 

alignment

T T A C A G T T C
A T ▲ C A ▲ ▲ T C

b) how score --- scoring matrix

IV) Non-alignment based scores

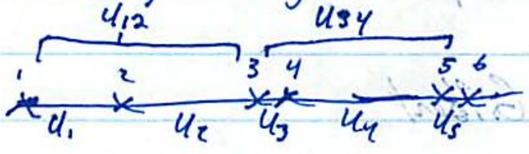
V) Database

- w/ lots of sequences in the database more likely getting a similarity by chance

VI) Profile

- e.g REST. ENZYMES
- expectation maximization -- finds motifs

VII) Distribution / Counting Words



ⓐ order $U_1 \cdots U_5$

ⓑ can also order U_{12}
U_{34}

Volker@grendel
3-9256

1/12/95

Math & Comp. Biology

Sequence Comparison I - Pairwise

DNA ↓ $\{A, C, T, G\}_4$

PROTEIN ↓ $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}_{20}$

Example

$\overset{7}{A}$ = A G C C T A G
$\overset{7}{B}$ = C A G C T G A

A G C C T ▲ A G
C A G C ▲ T G A

Alignment Parameters

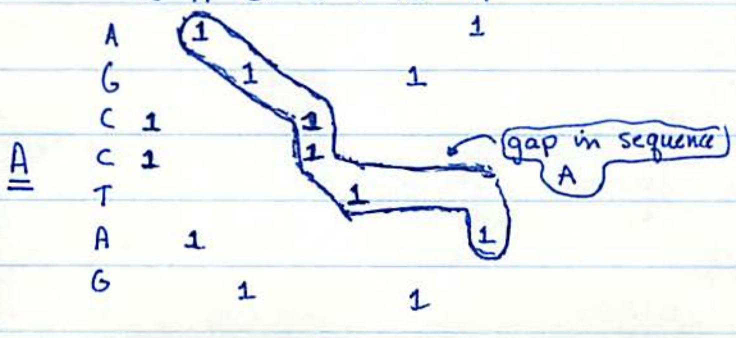
① mismatches
② matches
③ gaps
④ overhangs

Problem - too many possibilities

Solution I - Needleman & Wunsch (1970) JMB 48:443-453

Ⓐ represent all possible alignments on a matrix - as a path

- can have directionality (e.g. 5' vs. 3', N vs C)
- ∴ usually interested in diagonal



A \ B	C	A	G	C	T	G	A
A		1					1
G			1			1	
C	1			1			
C	1			1			
T					1		
A		1					1
G			1			1	

- score for matches/mismatches
- alignments become diagonals
- gaps become vertical or horizontal lines

NEEDLEMAN-WUNSCH II

- given two sequences

$A_1^N = \{a_1, a_2 \cdots a_n\}$ $\qquad$ $B_1^M = \{b_1, b_2 \cdots b_m\}$

Ⓐ an alignment is given by a set of paired indexes

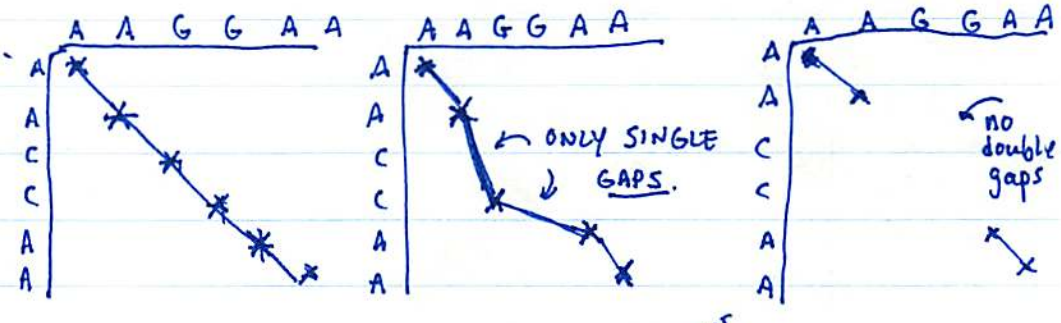
- $(a_{j1}, b_{k1})\ (a_{j2}\ b_{k2}) \cdots (a_{jL}, b_{kL})$
- with the restrictions

① $\left.\begin{array}{l} 1 \le j_1 < j_2 < \cdots j_L \le N \\ 1 \le k_1 < k_2 < \cdots k_L \le N \end{array}\right\}$ always increasing

② $\left.\begin{array}{l} \text{if } j_L > j_{L-1}+1 \text{ then } k_L = k_{L-1}+1 \\ k_L > k_{L-1}+1 \text{ then } J_L = J_{L-1}+1 \end{array}\right\}$ <u>no</u> double gaps

- that <u>is</u> in matrix the next match has to be in <u>either</u> next column or row.

```
- e.g   AACCAA    OR   AACG-AA   but not   AACC--AA
        AAGGAA         AA-GGAA             AA--GGAA
```



Ⓑ SCORING

Score for a path $= S_p = S_p(A, B)$

$= \sum_\ell \sigma(a_{j\ell}, b_{k\ell}) + \sum_L \omega(J_L - J_{L-1} - 1) + \sum_L \omega(k_L - k_{L-1} - 1)$

substitution scores

weight for gaps $\omega(0) = 0$, $\omega(L) < 0$ = cost

" # of gaps in each gap in A

" gaps in each gap in B

- substitution scores

σ Ⓐ SIMPLE

$\sigma(a, b) = 1$ if $a = b$ $\qquad$ $\omega(k) = -K$

$\sigma(a, b) = 0$ if $a \neq b$

(C) OPTIMIZATION

(1) - FIND MAX S_p FOR ALL PATHS

(2) ALGORITHM FOR FINDING MAX.

- depends on "no double gap" requirement

(A) SET $S_{\emptyset j} = \omega_j$ for <u>all</u> $j = 1, 2 \cdots M$

$S_{i \emptyset} = \omega_i$ for all $i = 1, 2 \cdots M$

} sets gap penaltles

optimal score for aligning $A_1^i = a_1 \cdots a_i$, $B_1^j = B_1 \cdots B_j$

$$S_{ij} = \max \begin{cases} S_{i-1,j-1} + \sigma(a_i, b_j) & A \\ S_{i,j-k} + \omega_k \text{ for all } k = 1, 2 \cdots j & B \\ S_{i-k,j} + \omega_k \text{ for all } k = 1, 2 \cdots i & C \end{cases}$$

then $S = S_{NM}$

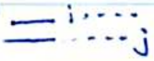
Proof

- optimal score for aligning 8 i, j <u>is</u>

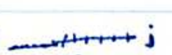
~~[illegible] i, j aligned~~

these are the only three possibilities to align i & j <u>If</u> assume <u>no</u> double gaps

(A) optimal score for aligning everything before i, j plus score of aligning i, j $= S_{i-1, j-1} + \sigma(a_i, b_j)$

(B) 

$= S_{i, j-k} + \omega_k$

(C) 

$= S_{i-k, j} + \omega_k$

RUNNING <u>NW</u>

	Δ	C	A	G	C	T	G	A
Δ	0	-1	-2	-3	-4	-5	-6	-7
A	-1	0 →	[illegible]	□				
G	-2							
C	-3							
C	-4							
T	-5							
A	-6							
G	-7							

G

Score for each cell = <u>Max</u>
of all 3 adjacent cells

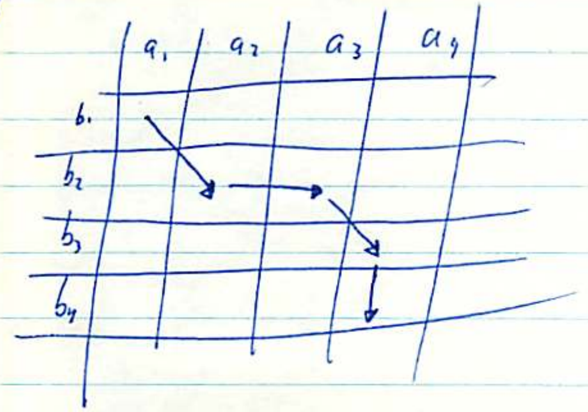
A
C A

Mathematic & Computation Molecular Biology - Pairwise Alignments

$A = \{a_1, a_2 \dots a_m\}$

$B = \{b_1, b_2 \dots b_n\}$

alignment = association of each $a_1 \dots a_m$ w/ each $b_1 \dots b_n$ (+ gaps)



↘ = assoc. of a_k b_ℓ

↓ = gap in b

→ = gap in a

<u>Needleman-Wunsch</u> - <u>corrected</u>

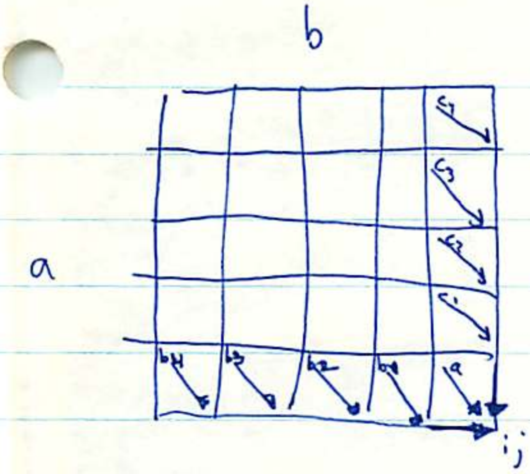
- no right angles *
- no double gaps

P = path = $\{a_{i_\ell}, b_{j_\ell} \cdots a_{i_1}, b_{j_1}\}$

$0 = i_0 < i_1 < \cdots < i_\ell < i_{\ell+1} = m+1$ ℓ = # of diagonals = alignment of position

$S_P = \underset{\text{for path}}{\text{Score}} = \sum_{k=1}^{\ell} \sigma\{a_{i_k}, b_{j_k}\} + \sum_{k=1} gap(k)$

$S = \text{max score} = \max S_P(A, B)$



— for last position, multiple paths

- if $\underline{a}$ $S = S_{i-1,j-1} + \sigma\{a_i, b_j\}$

if b_k $S = S_{i-1,j-k-1} + \sigma(a_i, b_{j-k}) + w_k$

if c_k ~~$S = S_{i-1-k,j}$~~ $+ \sigma(a_{i-k}, b_j) + w_k$

$S = S_{i-1-k,j-1}$

$S_{ij} = \max \{$

$-2 + 10 = 8$

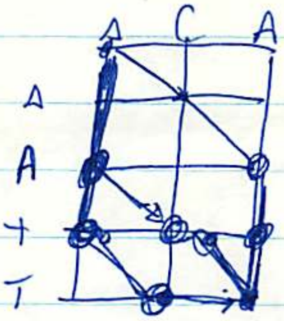
Example match=10

mismatch = -10

$w_1 = -2$

$w_2 = -5$

$w_3 = -8$



$-5 - 10 = -15 - 2$

$= -17$

-2

OLD
NW

	A	C	A
A	0	-2	-5
A	-2	-4	8
T	-5	-6	6
T	-8	-8	4

- can only work if $kw_1 \leq w_k$
- that is gap cannot jump too much

- this # is wrong - can never get 4

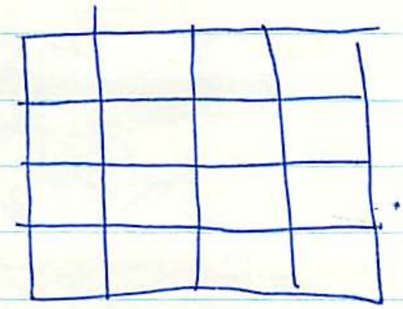
	A	C	A
A			
A			
T			
T			

AFFINE GAP PENALTIES

GOTOH Let $w_k = -\alpha - \beta k$ - $k = 1, 2 \ldots$ $\alpha, \beta > 0$

$\therefore$ for all k, α, β $k w_1 \leq w_k$

- This allows speeding up of algorithm



Proof

$H_{IJ} = \max \begin{cases} S_{I,J-1} + w_1 \\ S_{I,J-k} + w_k \quad k = 2 \ldots n \end{cases}$

$= \max \begin{cases} S_{I,J-1}, w_1 \\ \end{cases}$

$H_{IJ} = \max \{ S_{I,J-k} + w_k \} \quad k = 1 \ldots n \quad = \max \begin{cases} S_{I,J-1} + w_1 \\ H_{I,J-1} - \beta \end{cases}$

$V_{IJ} = \max \{ S_{I-k,J} + w_k \} \quad k = 1 \cdots n$

$S_{ij} = \max \begin{cases} S_{I-1,J-1} + \sigma(a_I, b_J) \\ H_{IJ} \\ V_{IJ} \end{cases}$

These assume that everthing ends in corner

- Thus end gap penalties are assessed

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

- But what about if you didnt want to assess these penalties

I) MODIFICATION I

- start w/ all 0's in columns $S_{I0} = S_{0J} = 0$
- find max score in last row or column

$S = \max \{ S_{mJ}, S_{IN} \}$

MODIFICATION II - LOCAL ALIGNMENT

A A ATTAC A A A

ATTGC G G G

$S_{I0} = S_{0J} = 0$ a) MAKE ▲ COLUMN/ROW ALL 0

b) AS SOON AS YOU GET A NEGATIVE # ... MAKE IT ∅.

$$S_{IJ} = \max \begin{cases} S_{I-1,J-1} + \sigma(a_I, b_J) \\ H_{I,J} \\ V_{I,J} \\ \emptyset \end{cases}$$

Mathematical & Computational Molecular Biology

FUNDAMENTALS

① alignment ⟷ lattice path

② scoring

- substitutions

  DNA = 10 scores 4x4 matrix = 4+3+2+1

  proteins = 210 scores 20x20 matrix = 20+19+ ... +2+1

- gap penalties

  - most use affine gap penalties $\alpha + \beta k = w_k$; $\alpha, \beta < 0$

③ algorithm - how find high scoring alignments

④ interpretation

- how meaningful are alignments (statistics)

LOCAL - SMITH - WATERMAN

- CATTGC vs ATG
- mismatch = -1
- match = +3
- $w_k = -1-k$

	▲	C	A	T	T	G	C
▲	0	0	0	0	0	0	0
A	0	0	3	1	0	0	0
T	0	0	1	6	4	3	2
G	0	0		4	5	7	5

multiple pathways

① take max {all poss. scores, 0}

② 0's in first row

③ fill #'s

④ pick highest score

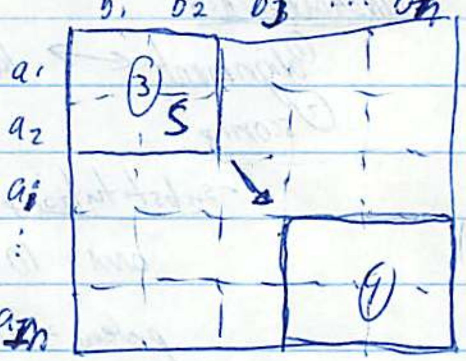
⑤ trace path

So - what do you do with multiple paths?

Zuker - Suboptimal Alignments

① constrain alignments such that they must go through certain points aI, bJ

② find all (aI bJ) such that there is at least one alignment containing (aI, bJ) with score $S \geq S_{mn} - \delta$



two separat subproblems for each point

③ want to find $S_{I-1, J-1}$ = optimal score

④ can solve for this region by inverting sequences $(S_{I+1 \dots m, J+1 \dots n})$

Noor & Brutlag - Near optimal alignment

brutlag@cmgm

- motivation .. are structural & optimal alignments the same

- Vingron & Argos .. reliably aligned regions

Vary parameters

Dayhoff = PAM matrix

why not plot in 3D w/ score on Z axis

1/24/95

Math/Comp

username@grendel /USER2/username

SEQF

ZUKER - allows double gaps

$$\begin{array}{cc} \text{---} & bbb \\ aaa & \text{---} \end{array} \quad w_{k+l} \text{ not } w_k + w_l$$

STATISTICS

ⓐ for full NW alignment w/ gaps --- no good theory for getting statistics

- so ⓐ shuffle the seqs 100x . 1000x . .
  ⓑ order the scores
  ⓒ how many is score above

- Problem
  ⓐ computer intensive
  ⓑ changes w/ gap penalties ... must redo for each

- Solutions -
  ① Disallow gaps (or v.v. high penalty)
  ⓐ all alignments on diagonals
  ⓑ take a particular diagonal

$b_1 \cdots b_n$
$a_1 \cdots a_{1+n-i}$

ⓒ replace alignment w/ sequence of scores
ⓓ find segment pairs w/ highest aggregate score

Statistical Theory

- Scores : $S_1, S_2, \dots S_r$ r different score

w/probabilities: $p_1, p_2, \dots p_r$

random variable $x_1, x_2, \dots x_N$ prob $\{x_i = S_k\} = p_k$

$$S_k = \sum_{i=1}^{k} x_i$$

$$M = \max\{S_l - S_k\} \quad 0 \le k \le l \le N$$

EXAMPLE

	p_k
$S_1 = -5$	0.2
$S_2 = -3$	0.2
$S_3 = -1$	0.2
$S_4 = 1$	0.2
$S_5 = 5$	0.2

$$-3, \underbrace{-1, 5, 1, 5, 5,}_{16} -3$$

$S_1 = -3$ $S_4 = 2$ $S_7 = 9$

$S_2 = -4$ $S_5 = 7$

$S_3 = 1$ $S_6 = 12$

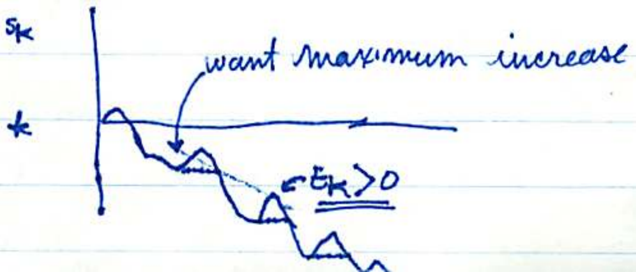
$\max = S_6 - S_2 = 16$

2 Assumptions

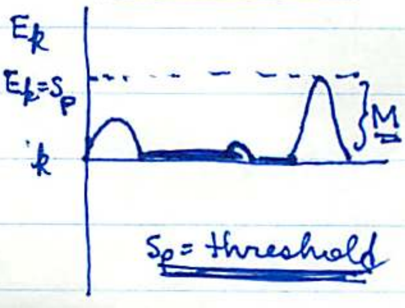
- $E[X] = \sum_{k=1}^{r} p_k s_k < 0$

- $\text{Prob}(x_1 > 0) > 0$

RESULTS



EXCURSION PLOT



S_p = threshold

$E_0 = 0$

$E_k = \max \left\{ \begin{matrix} E_{k-1} + S_{k+1} \\ \text{OR} \\ \emptyset \end{matrix} \right\}$

So... how calculate S_p?

$$\text{Prob}\left\{M > \frac{\ln N}{\lambda} + x\right\} \cong 1 - e^{-Ke^{-\lambda x}}$$

N = length
x = pos. variable

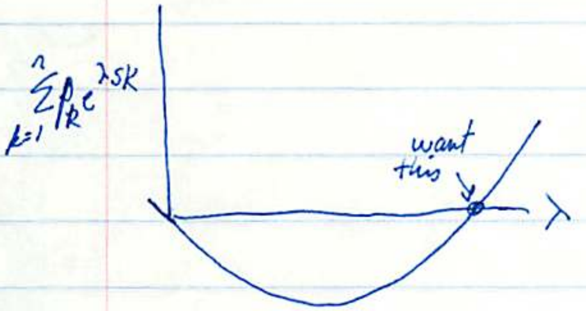
$\left.\begin{matrix}\lambda \\ k\end{matrix}\right\}$ positive parameters

$$\text{Prob}\left\{M > S_p\right\} \cong 1 - e^{-kNe^{-\lambda S_p}}$$

① λ is unique positive root for $\sum_{k=1}^{n} p_k e^{\lambda S_k} = 1$

$$\Downarrow$$

$$e^{\lambda S_k} = e^{\lambda' \alpha S_k}$$



λ is scale parameter

① if have scores $S_1, S_2 \cdots S_r$
w/ λ, S_p

② if have other scores $S_{b1} \cdots S_{br} = \alpha S_k$

② $\lambda' \alpha = \lambda$ $\quad S_p' = \alpha S_p$

$\lambda' = \lambda/\alpha$

② $$k = \frac{F e^{-2(A+B)}}{\lambda C}$$

$$A = \sum_{k=1}^{\infty} \frac{1}{k} E\left[e^{\lambda \$_k}; \$_k < 0\right]$$

$$B = \sum_{k=1}^{\infty} \frac{1}{k} \text{prob}\left\{\$_k \geq 0\right\}$$

$$C = E\left[\$_1 e^{\lambda \$_1}\right]$$

F = correction factor for non-additive

1/26/95

Mathematical Molecular Biology

mmbjc@cmgm ... gbrand

SEQUENCE OF SCORES $X_1, X_2 \cdots X_N$ i

prob. $\{X_i = S_k\} = p_k$ $i = 1 \cdots N$ $k = 1 \cdots r$

$S_k = \left\{ \begin{array}{l} 0 \\ \cdot \end{array} \right.$

p_k = frequency over all the sequences

$Max = \max \{ S_e - S_k \}$

$Prob \{ M = S \} = 1 - e^{-KNe^{-\lambda S}}$

APPLICATION

- for given scheme + significance level p want S_p where $\text{prob} \{ M > S_p \} = p$

$$s_p = \frac{1}{\lambda} \left(\ln N + \ln K - \ln(\ln(1-p)) \right)$$

SAPS
- anonymous FTP ~~an~~ @ gnomic
- karlin.c calculates λ, K

COMPOSITION BIAS

- in high scoring segments the occurrence frequencies of scores are biased

$q_k = p_k e^{\lambda s_k}$ $\quad {}^{①}\sum q_k = 1$ $\quad {}^{②}$ if $s_k > 0 \Rightarrow q_k > p_k$

if $s_k < 0 \Rightarrow q_k < p_k$

$\Downarrow$

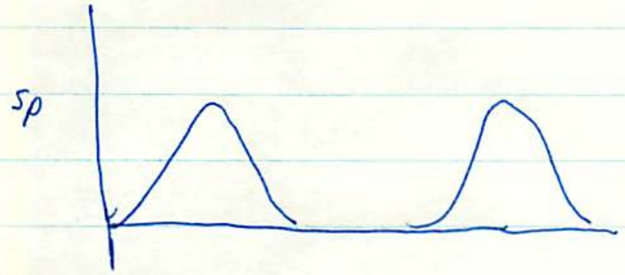
$s_k \propto \ln \frac{q_k}{p_k}$

- this can be reversed to ~~d~~calculate s_k for different biases of interest in alignments

EXPECTED LENGTH $E[L] \hat{=} \frac{S_p}{\sum q_k s_k}$

..PLICATIONS - TRANSMEMBRANE DOMAINS

① identify region
② make gk
③ evaluate w/ excursion plot



but might miss segments containing two distinct domains

APPLICATION - CHARGE CLUSTERS

	S104	
K, R	+2	+2
D, E	-2	-8
others	-1	-5
	cluster	run of + chags

HYDROPHOBICITY

2ARY STRUCTURE

doesnt work well because no positional information

Application to Sequence Comparisons

- for a given alignment

M = length $b_1 \quad b_2 \cdots b_{N-i+2}$ $\leftarrow p'_k \quad k = 1 \ldots r$

N = length $a_I \quad a_{I+1} \cdots\cdots a_N$ $\leftarrow p_k \quad k = 1 \cdots r$

$\sigma \quad \sigma$

· sequence of scores $\sigma \quad \sigma \quad \sigma$

· So get scores for all possible diagonals (no gaps)

· M = maximal segment pair (aggregate score)

$\text{Prob}\{M > s\} = 1 - e^{-kMNe^{-\lambda s}}$ - k, λ analogous to before
- λ = unique positive root

Target frequencies = freq. of substitution

$$q_{ij} = p_i p_j e^{\lambda s_{ij}}$$

p_k must be somewhat similar to p'_k

$$s_{ij} \propto \frac{q_{ij}}{p_i p'_j}$$

2/1 wed 3:30-5:00pm
Winston Hide – MASSPAR -- Database Searching

1/31/95

Mathematical & Computational Molecular Biology

Amino Acid Substitution Scores

① 20 × 20 matrix

-almost always symmetrized

② given these scores ... high scoring segment pairs have a ~~distinct~~ biased composition

$$q_{ij} = \underbrace{p_I p_J}_{\text{prob. of generating sequences } I,\, J} e^{\lambda S_{IJ}} \quad \leftarrow \text{scale factor}$$

$$\underline{S_{IJ}} = \underset{\text{proportionality constant}}{\rho} \underbrace{\ln \frac{q_{IJ}}{p_I p_J}}_{\text{log-odds score}}$$

= these scores will target those regions w/ q_{IJ}

③ Assumptions

ⓐ at least one score > 0

ⓑ E $\sum p_I s_I < 0$

Derivation of Substitution Scores

① A C L L M A G
A C V I M G A

② count substitutions k_{ij}, k_{ji}

③ symmetrize $C_{ij} = k_{IJ} + k_{JI}$

④ take row, column sums

rows = $C_{I\cdot}, C_{2\cdot}, C_{3\cdot} \cdots C_{20\cdot}$

columns = $C_{\cdot 1}, C_{\cdot 2}, C_{\cdot 3} \cdots C_{\cdot 20}$

$$\begin{array}{lll} C_{11} & C_{12} & \cdots C_{1,20} \\ C_{21} & C_{22} & \\ C_{2\cdot} & & \\ \vdots & & \\ \vdots & & \\ C_{20,1} & C_{20,2} & \end{array}$$

C_{IJ} = # of subs. I↔J

C_{II} = 2x # of I↔I matches

$C_{\cdot I} = C_{I\cdot}$ = # of I residues

$C_{\cdot\cdot}$ = total # residues

$f_i = \frac{C_{i\cdot}}{C_{\cdot\cdot}}$ = freq of residue i

③ Getting log-odds scores

ⓐ Contingency Table Approach

$$S_{IJ} = \ln \frac{C_{IJ}}{(C_{I\cdot})(C_{\cdot J})/C_{\cdot\cdot}} = \frac{obs}{expected}$$

$$= \ln \frac{f_{IJ}}{f_I \cdot f_J}$$

④ How get counts?

ⓐ BLOSUM (Henikoff & Henikoff)

- take blocks
- each position make a column vector of diff residues
- consider all pairwise comparisons
- each used for substitutions

5A	AA = 10	AC = 15
3C	CC = 3	CS = 6
2S	SS = 1	AS = 10

$$\begin{array}{c|ccc|c} & A & C & S & \\ \hline A & 20 & 15 & 10 & 45 \\ C & 15 & 6 & 6 & 27 \\ S & 10 & 6 & 2 & 18 \\ \hline & 45 & 27 & 18 & \end{array}$$

ⓑ STRUCTURE

ⓒ PW Alignments

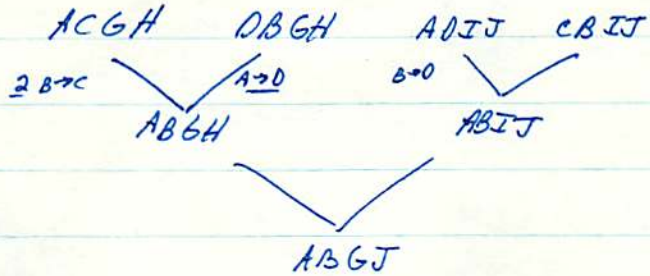
ⓒ PAMs - inferred from evolutionary data

ⓑ PAMs - Dayhoff

ⓐ infer substitutions from a likely evolutionary tree

- only works for highly conserved proteins

ⓑ example - 4 seqs



ⓒ use this to get substitution counts K_{IJ}

ⓓ symmetrize to C_{IJ} — so both directions the same

→ ⓔ use transition probabilities of change over time

- MARKOV MODEL

- changes are governed by probability transition matrices

ASSUMPTIONS
- mutations occur at constant rate
- accepted changes

M_{IJ} (20 x 20) = prob. that J changes to I in 1 unit of time

$$\sum_{i=1}^{20} M_{IJ} = 1$$

1/31/95

PAMS

(1) Define relative mutability ~ how often does a particular aa change?

$$m_J = \frac{C_{\cdot J} - C_{JJ}}{C_{\cdot J}} = \frac{\text{\# residue involved in substitution}}{\text{all of that residue}}$$

(2) 1-step transition probability

(a) $$\underline{M_{JJ} = 1 - \rho m_J}$$ = prob. that J changes

- ρ is set in way such that # changes in 1 unit time = 1%

set $\sum_{J=1}^{20} f_J M_{JJ} = 0.99$: fraction not changing

solve $\rho = \frac{1}{100 \sum_{J=1}^{20} f_J m_J}$ 1 accepted point mutation per 100 residues

(3) $I \neq J$

$$\underline{M_{IJ} = \rho m_J \frac{C_{IJ}}{C_{\cdot J} - C_{JJ}}}$$

PROPERTIES

(1) $\sum_{I=1}^{20} M_{IJ} = 1$

(2) M_{IJ}^{n} = prob. of J → I in n units

$$M_{IJ}^{(n)} = (M_{IJ})^n$$

(3) Stationary distribution

$$\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{20} \end{pmatrix} \therefore (M_{IJ}) \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{20} \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{20} \end{pmatrix}$$

- ∴ only works for proteins with the same composition

(4) $f_J M_{IJ} = f_I M_{JI}$

$f_J M_{IJ}^{n} = f_I M_{JI}^{n}$ } ∴ SYMMETRICAL LOG ODDS SCORES

⑧ $S_{IJ}^{n} = \ln \frac{M_{IJ}^{n}}{f_I}$ LOG-ODDS SCORES

... as $n \to \infty$ $M_{IJ} \to f_I$

$\therefore$ all scores $\to \emptyset$

Which PAM's

① n small ... for highly conserved

n large ... for highly divergent

Mathematical and Computational Molecular Biology
Math/Biol 227 – Brendel – WQ94/95

References: Score-based sequence analysis.

Altschul, S.F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219**, 555–565.

Altschul, S.F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* **36**, 290–300.

Collins, J.F., Coulson, A.F.W. & Lyall, A. (1988). The significance of protein sequence similarities. *Comput. Appl. Biosci.* **4**, 67–71.

Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.). vol. 5, suppl. 3, pp. 345–352. Nat. Biomed. Res. Found., Washington, DC.

Dembo, A. & Karlin, S. (1993). Central limit theorems of partial sums for large segmental values. *Stoch. Proc. Appl.* **45**, 259–271.

Henikoff, S. & Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915–10919.

Henikoff, S. & Henikoff, J.G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins* **17**, 49–61.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Bosci.* **8**, 275–282.

Karlin, S. (1994). Statistical studies of biomolecular sequences: score-based methods. *Phil. Trans. R. Soc. Lond.* B **344**, 391–402.

Karlin, S. & Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2264–2268.

Karlin, S. & Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science* **257**, 39–49.

Karlin, S. & Cardon, L. (1994). Computational DNA sequence analysis. *Annu. Rev. Microbiol.* **48**, 619–654.

Karlin, S. & Dembo, A. (1992). Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Prob.* **24**, 113–140.

Karlin, S., Dembo, A. & Kawabata, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.* **18**, 571–581.

Mathematical and Computational Molecular Biology
Math/Biol 227 – Brendel – WQ94/95

References: Sequence alignments with gaps.

Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705–708.

Naor, D. and D.L. Brutlag (1994) On near-optimal alignments of biological sequences. *J. Comp. Biol.* ~~4, 000–000~~. 1(4) = 1–18 ?

Needleman, S.B. and C.D. Wunsch (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

Smith, T.F. and M.S. Waterman (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

Zuker, M. (1991) Suboptimal sequence alignment in molecular biology. Alignment with error analysis. *J. Mol. Biol.* **221**, 403–420.

Genomic Signatures

Dinucleotide Relative Abundances -- invariant w/in species

ρ^*_{xy} <0.77	ρ^*_{xy} 0.77-0.81	0.81-1.18	1.18 -	
- -	-	0	+	++

Distances

residual dinuc = $\frac{f^*_{ij}}{f^*_i \cdot f^*_j} - 1$

- $\rho^*(f,g) = \sum \left| \rho^*_{IJ}(f) - \rho^*_{IJ}(g) \right| w_{ij}$ ∍ $w_{IJ} = 1/16$

- use

① Ind. Ident Distrib

$\rho^*_{xy} \rightarrow 1$

② betw. species distances generally greater than w/in species distances

#s v. low w/in a species E.coli

	1	2	3	4
1	0			
2	0.025	0		
3	0.008	0.017	0	
4	0.032	0.014	0.025	0
5	0.009	0.021	0.008	0.027

USES AVERAGE VALUES

ADVANTAGE OF THE DISTANCE
① di α tetra α tri
② ρ^* XNKY - ~~any~~ k this ~~[illegible]~~ approaches 1

$\boxed{1}\boxed{3}\boxed{5}$ $\boxed{2}\boxed{4}$

$\boxed{5}\boxed{1}\boxed{3}$

Explanations

(A) why is TA low?

① lowest stacking energy of dinucs
- most unstable dinuc.

but doesn't explain non-coding bias ← ② RNase preferentially degrade UpA

③ part of many regulatory sequences

(B) why is CG low?

① methylation → deamination

NH_2-CpG

$\downarrow$

$\overset{CH_3}{NH_2-CpG}$

$\downarrow$

$\overset{CH_3}{TpG}$

$\downarrow$

- but doesn't explain why CpG low in mitochondria because no methylation

(C) CTAG low

① kinks easily
② clusters in rRNA
③ dense in replication origins

(D) GATC low in almost all bacteriophage in E.coli

If

$\rho^*_{xy} \leq 0.78$ = highly significant underrepresentation

$\rho^*_{xy} \geq 1.22$ = " " overrepresentation

Examples

① - CG is low for all vertebrates

- single mutants from this

$$\left.\begin{array}{l} CG \to CA \\ \;\;'' \;\;\to TG \end{array}\right\} \text{overrepresented in animals}$$

② Proks

- TA v. low

- $\overset{GATC}{C\boxed{TAG}}$ v. low .... embedded in this is 2 stop codons

- ATAG v. high ...but other stop codons not low.

$$\frac{f_{xyz}}{\underbrace{f_x f_y f_z}_{\text{factoring out mononucl. freqs}} \underbrace{\left(\frac{f_{xy}}{f_x f_y}\right)\left(\frac{f_{yz}}{f_y f_z}\right)\left(\frac{\overset{XNZ}{f_{x\cdot z}}}{f_x f_z}\right)}_{\text{residual dinucleotide effects}}}$$

$$\frac{f_{xyz}\ f_x f_y f_z}{f_{xy}\ f_{x\cdot z} f_{yz}} = \text{TRINUCLEODIE FREQS W/ ALL MONO, \underline{DI} REMOVES}$$

$$\frac{f_{xyzw}}{f_x f_y f_z f_w \left(\begin{matrix}\text{dinucleotide}\\\text{residuals}\end{matrix}\right)\left(\begin{matrix}\text{trinucleotide}\\\text{residuals}\end{matrix}\right)} = \text{TETRA W/ ALL MONO, DI TRI. REMOVED}$$

<u>REMOVING</u> <u>STRANDS</u> - SYMMETRIZING

$$f^*_{xy} \qquad f^*_A = f^*_T = \frac{1}{2}(f_A + f_T)$$

$$f^*_{GT} = f^*_{AC} = \frac{1}{2}(f_{GT} + f_{AC})$$

TAKE LONG ENOUGH SEQUNCE THEN THE $f_A = f_T$,

$$\rho^*_{xy} = \frac{f^*_{xy}}{f^*_x f^*_y}$$

Dinucleotides

① How measure bias

ⓐ longer is better >19000 needed

ⓑ mononucleotide content

- GC varies between 10-80%
- doesn't fit w/ habitat
- must factor out mononucleotides

$$- E(f_{xy}) = f_x \cdot f_y$$

ⓒ

$$\boxed{\text{- ODDS RATIO} \quad \frac{f_{xy}}{f_x f_y} - 1} = \text{DINUCLEOTIDE BIAS}$$

ⓓ trinucleotides

- Markov model of order 1

~~fABC~~

$$f_{ABC} = Pr\{A \overset{\text{given}}{|} BC\} \, Pr\{BC\}$$

← since order 1 C has no effect on A

$$= Pr\{A|B\} \, Pr\{BC\}$$

$$= \frac{Pr\{AB\} \, Pr\{BC\}}{Pr\{B\}}$$

ORDER 2

$$f_{ABCD} = \frac{Pr\{ABC\} \, Pr\{BCD\}}{Pr\{BC\}}$$

ORDER n-2

$$f_{x_1 \cdots x_n} = \frac{f\{x_1 \cdots x_{n-1}\} \; f\{x_2 \cdots x_n\}}{f\{x_2 \cdots x_{n-1}\}}$$

Mathematical & Computational MolBio - Karlin

2/2/95

Enormous database of sequences

DNA - 200,000,000 bit

PROT ... 60,000 bit

Structures

Lots of opportunities to see patterns in data

①complete genomes

- which are genes
- repeats
- word patterns

②short words

- over/under representations & biases
- compare diff. pieces of DNA

②mononucleotides

①G+C varies enormously
- e.g. isochores in humans
- not biased in bacteria, flies

Blaisdell ②Markov models ... of any order
- show that DNA's are not dependent on neighboring bases
- too many local repeats

③ Chaos models -- not yet applied well

④ Linguistic models ... doesn't think this is good either
- freq. of words

- DOESN'T BELIEVE IN FITTING MODELS
- BELIEVES IN BENCHMARKS

$\log\frac{1}{10} = \log 1 - \log 10 = 0 - 1 = -1$

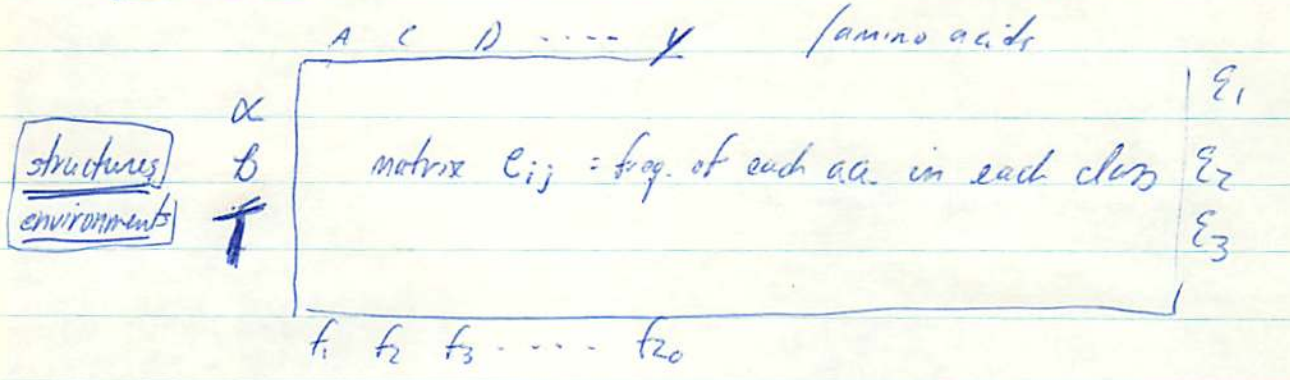
Math / Comp. Molecular Biology

Substitution Matrices

- substitution scores $\{s_{ij}\}$ $\begin{matrix} i=1\ldots 20 \\ j=1\ldots 20 \end{matrix}$; symmetrical
- matrices -- do <u>not</u> have to be square <u>NOR</u> symmetrical

<u>EXAMPLE</u> - COMPARE STRUCTURES & SEQUENCES = INVERSE FOLDING

① <u>derive counts</u>



② form scores $s_{ij} = \ln\frac{e_{ij}}{\varepsilon_i f_j} = \ln\frac{\text{observed}}{\text{expected}}$

③ use for comparisons

- s_{ij} occurs w/ prob. $p_i p_j'$

① in HSPs substitution freqs are ~ $q_{ij} = p_i p_j' e^{\lambda s_{ij}}$

② typical <u>length</u> of HSP at a given significance p is

$$L_p = \frac{S_p}{H/\lambda}$$

S_p = signifant score threshols $= \frac{1}{\lambda}\{\ln MN + \ln k - \ln[-\ln(1-p)]\}$

∴ as H incr. the length shrinks
as H decr. the length increases

$H = \lambda \sum_{i=1}^{i=20}\sum_{j=1}^{j=20} q_{ij} s_{ij}$ = "<u>relative entropy</u>" (Altshul)

-H varies among matrices

(3) F_I = expected fraction of identities in HSSP

$$F_I = \sum_{i=1}^{i=20} q_{ii}$$

$$F_P = \sum_{i=1}^{i=20} \sum_{j=1}^{j=20} q_{ij} \quad (q>0)$$ = fraction of conservative substitutions

CREATING MATRICES

(1) PIM$_F$ = percent identity matrices
- target HSSP where $F_I = F$

(a) start w/ subs. counts from learning set ⇒ c_{ij}
- derive from segments w/ a lot of identities
- assume ratio of changes will be constant even w/ lower identities

Conditionals $g_{ii} = \frac{c_{ii}}{\sum c_{ii}}$ $\qquad h_{ij} = \frac{c_{ij}}{c_{..} - \sum c_{ii}}$

$$s_{ii} = \ln \frac{F g_{ii}}{f_i f_i} \qquad s_{ij} = \frac{(1-F) h_{ij}}{f_i f_j}$$

(b) $\sum_{i=1}^{20} \sum_{j=1}^{20} f_i f_j e^{\lambda s_{ij}} = 1$

(c) $F_I = \sum_{i=1}^{20} q_{ii} = \sum_{i=1}^{i=20} f_i f_i e^{s_{ii}} = \sum_{i=1}^{20} F g_{ii}$

BLAST

- also includes some correction for multiple sequences in the database
  - adjusts ℓ so that length is based on length of entire databases

PAIRWISE ALIGNMENTS

SSPA - significant segment pair alignment

1. determine all HSSPs
2. order HSSPs optimally
3. eliminate overlaps
4. $\text{score} = \dfrac{\Sigma \text{HSSPs}}{\text{max. self score}}$ or $\dfrac{\Sigma}{\text{min self score}}$ } GLOBAL SCORE

~~score~~

$$n \quad \text{score} = \frac{\Sigma \text{HSSPs}}{\cancel{\text{range}} \text{ max or min for range of alignment}}$$

$$\text{score} = \frac{\Sigma \text{HSSPs}}{\text{aligned region}}$$

2/9/95

Phylogeny

- taking sequences and ordering a tree
- organizing relationships to reflect evolutionary descent

(1) Assume evolutionary changes are caused by mutations that are substitutions or andels or inversions

Complications

(1) observed now - infer past

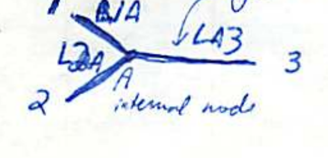
3 types of similarities

- ancient shared characteristics
- derived shared characteristics
- convergent shared characteristics

Trees

3 species

L1A+L2A
L1A+LA3
L2A-LA3



edges... length represents divergency/distance

D_{12} D_{13} D_{23}

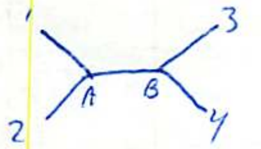
$$\boxed{\begin{aligned} L1A + L2A &= D12 \\ L1A + LA3 &= D13 \\ L2A + LA3 &= D23 \end{aligned}}$$

$$\boxed{\begin{aligned} L1A &= \tfrac{1}{2}\{D12 + D13 - D23\} \\ L2A &= \tfrac{1}{2}\{D12 + D23 - D13] \\ LA3 &= \tfrac{1}{2}\{D13 + D23 - D12\} \end{aligned}}$$

UNIQUE SOLUTION

-~~but~~ unique solution is meaningless

4 species



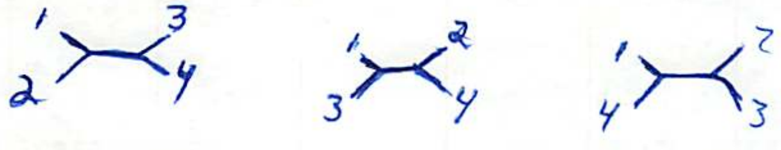
: $\binom{4}{2}$ distances = 6 .

D_{12} D_{13} D_{14}
D_{23} D_{24}
D_{34}

5 Branches

L_{1A}
L_{2A}
L_{AB}
L_{B3}
L_{B4}

① system is overdetermined
② but ... other trees ... 3 total $2^{n-2}-1$

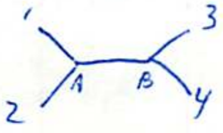


In general

-n species	n	2	3	4	5	6	n	n+1
-D_n pairwise distances	$\binom{n}{2}$	1	3	6	10	12	D_n	D_n+n
-L_n branch lengths	$2n-3$	1	$3+2=$	$5+$	$2=7$		L_n	L_n+2
-T_n topologies	$(1\cdot3\cdot5\cdots2n-5)$	1	1	3	15		T_n	L_nT_n

$$\frac{(2n-5)!}{(2^{n-3})(n-3)!}$$

4 species



- even w/ unequal rates -- can factor out rates
- because have a shared segment

$L_{1A} + L_{A2} = D_{12}$
$L_{1A} + L_{AB} + L_{B3} = D_{13}$
$L_{1A} + L_{AB} + L_{B4} = D_{14}$
$L_{2A} + L_{AB} + L_{B3} = D_{23}$
$L_{2A} + L_{AB} + L_{B4} = D_{24}$
$D_{B3} + L_{B4} = D_{34}$

<u>what is the condition for solution to exist?</u>

① * $D_{12} + D_{34} \le D_{13} + D_{24} = D_{14} + D_{23}$

① <u>If</u> $D_{13} + D_{24} = D_{14} + D_{23} \Rightarrow$ there is a unique solution with properties *

② <u>But</u> generally use least-squares estimate to approximate solution L_{ij}

∴ try to minimize <u>LS</u>

<u>LSQ solution</u>

$$L_{1A} = \frac{1}{4}\{D_{13} + D_{14} - D_{23} - D_{24}\} + \frac{1}{2} D_{12}$$

$$L_{2A} = \frac{1}{4}\{D_{23} + D_{24} - D_{13} - D_{14}\} + \frac{1}{2} D_{12}$$

$$L_{B3} = \frac{1}{4}\{D_{13} + D_{23} - D_{14} - D_{24}\} + \frac{1}{2} D_{34}$$

$$L_{B4} = \frac{1}{4}\{D_{14} + D_{24} - D_{13} - D_{23}\} + \frac{1}{2} D_{34}$$

$$L_{AB} = \frac{1}{4}\{D_{13} + D_{14} + D_{23} + D_{24}\} - \frac{1}{2}(D_{12} + D_{34})$$

if $D_{13}+D_{24}=D_{14}+D_{23}$ =

For additive tree

$$L_{1A}=\frac{1}{2}\{D_{12}-D_{23}+D_{13}\}$$

When sequences are v. similar most mutations are independent ∴ distances are additive

Fitch

- assumes relaxed additivity
- take 4 species at a time

Methods

① Find correct tree & estimate branch lengths

distance
parsimony
likelihood

2/14/95

Distance matrix methods

= n species : (OTUs operational taxonomic units)

- $\binom{n}{2}$ distances

128

- $\frac{(2n-5)!}{(2^{n-3}(n-3)!}$ distinct bifurcating trees w/ OTU leafs $\left(\frac{123!}{2^{61}\cdot 61}\right)$

- $2n-3$ branches

Task

① find correct topology
② estimate branch lengths

Methods

① evaluate all trees -- use LS -- get distances -- pick smallest

2/14/95

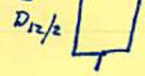
UPGMA = unweighted pair group method w/ arithmetic mean

Observed Distances

$$\begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & \\
1 & \cdot\, d_{11} & d_{12} & d_{13} & d_{14} & \cdots \\
2 & & & d_{23} & d_{24} & \cdots \\
3 & & & & & \\
4 & & & & &
\end{array}$$

Recursion (reduce # of species

① find smallest distance ... D_{12}

②  -make branches joined w/ length $D_{12}/2$

③ replace 1 & 2 by 12

④ convert $\underline{D_{X\text{-}12}} = \dfrac{D_{X1} + D_{X2}}{2}$

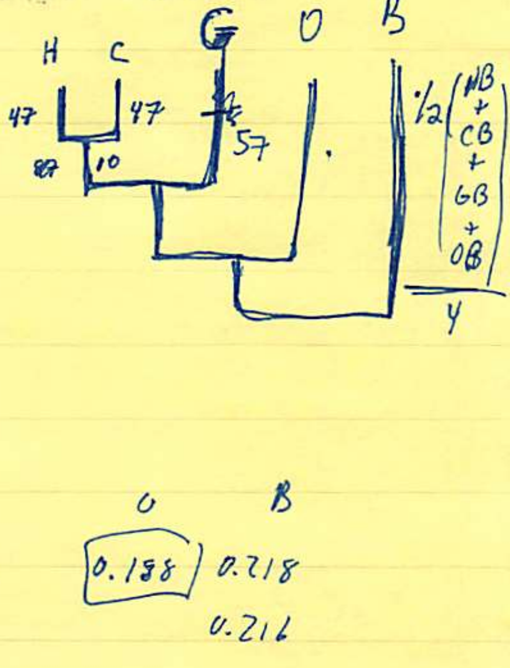
⑤

Example

	H	C	G	O	B
H		$\boxed{0.094}$	0.14	0.180	0.207
C			0.115	0.194	0.218
G				0.158	0.218
O					0.216
B					

↓

	HC	G	O	B
HC		$\boxed{0.1275}$	0.187 0.188	0.212 0.215
G			0.188	0.218
O			~~[illegible]~~	0.216
B				~~[illegible]~~


H
C
G
O
B
47
47
87
10
57
$\frac{1}{2}\left(\begin{matrix}HB \\ + \\ CB \\ + \\ GB \\ + \\ OB\end{matrix}\right) \Big/ 4$

→

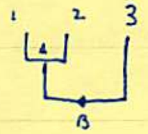
	O	B
HCG	$\boxed{0.188}$	0.218
		0.216

ADVANTAGES

- v.v. fast

Properties

- branch lengths always positive
- from any point internal to the leaf you get the same average distance
- ∴ only works best w/ constant evolutionary rate
- if get correct topology then the estimates of the branch lengths are least squares if assume constant rate

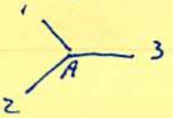


12 > 0
13 > 0
23 > 0

AB > 0

Fitch & Margoliash

- works on triplets



$$✱\ L_{1A} = \tfrac{1}{2}\{D_{12} + D_{13} - D_{23}\}$$
$$L_{2A} = \tfrac{1}{2}\{D_{12} + D_{23} - D_{13}\}$$
$$L_{3A} = \tfrac{1}{2}\{D_{13} + D_{23} - D_{12}\}$$

① start w/ matrix

② find smallest Distance D_{12}

③ group remaining distances into $1 \rightarrow n$

$$D_{1n} = \frac{D_{13} + D_{14} + D_{15} \cdots D_{1n}}{n-2}$$

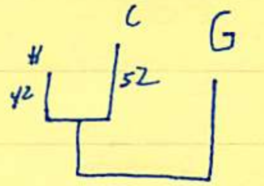
$$D_{2n} = \frac{D_{23} + D_{24} + D_{25} \cdots D_{2n}}{n-2}$$

④ calculate branch lengths from ✱

⑤ group 1, 2

⑥ then convert D_{1n} D_{2n} like UPGMA … w/ averages

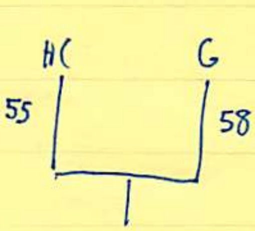
Example --

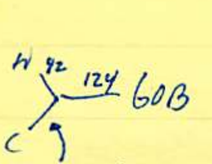


$D_{HC} = 94$

$D_H(60B) = \frac{1}{3}(111 + 180 + 207) = 166$

$D_C(60B) = 176$

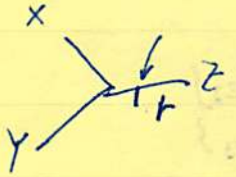




$1/2(D_{12} + D_{23} - D_{13}) = 52$

$8 = 55 - \frac{(42+52)}{2}$

rooting



Assume constant rate

$$2\lambda t = r$$
$$= x+z-r$$
$$= y+z-r$$

Least Squares $\bar{r} = \frac{1}{4}(x+y+2z)$

Then -- do branch swapping

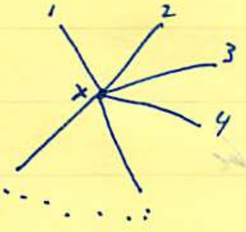
-- try to minimize Least Squares

Neighbor - Joining   Saitou & Nei 1987 JME 4:406-425

① n species w/ matrix

② derive topology differently

③ begin w/ star-like topology



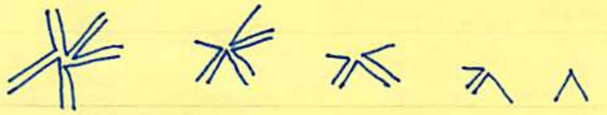
: L_{1x}, L_{2x}, L_{3x}, L_{4x} --- L_{nx} = branches

: distances = D_{ij}

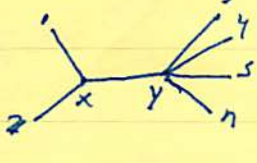
④ want to minimize sum of branch lengths

⑤ assume branch lengths are additive

$$S_0 = \sum_{i=1}^{n} L_{xi} \overset{\text{ASSUME ADDITIVE}}{=} \frac{1}{n-1} \sum_{i<j} D_{ij}$$



⑥ consider diff. topologies



length = L_{12} L_{1x} L_{2x} L_{xy} L_{y3} · L_{y4}

SUM FOR THIS TOPOLOGY = $S_{12} = \underbrace{L_{1x} + L_{2x}}_{D_{12}} + L_{xy} + \underbrace{\sum_{i=3}^{n} L_{yi}}_{\frac{1}{n-3}\sum D_{ij}} \overset{\text{ASSUME ADDITIVE}}{\Rightarrow} D_{12} + \frac{1}{n-3} \sum_{3 \le i < j}^{n} D_{ij}$ $\cancel{+ \sum_{k=3}^{n} D_{1k} - L_{1x} - L_{yk}}$

$$+ \frac{1}{2(n-2)} \sum_{k=3}^{n} (D_{1k} - L_{1x} - L_{yk} + D_{2k} - L_{2x} - L_{y2})$$

$$= \frac{1}{2(n-2)} \sum_{k=3}^{n} \left(D_{1k} + D_{2k}\right) + \frac{1}{2} D_{12} + \frac{1}{n-2} \sum_{3 \le i < j} D_{ij}$$

⑦ calculate <u>all</u> possible D_{xy} $\binom{n}{2}$

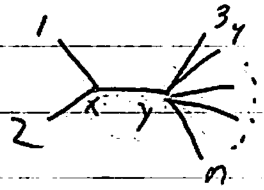
⑧ choose that which gives minimum sum

⑨ reduce distance matrix by taking averages

It turns out - that the distances come out to be the least squares estimate.

2/16/95

Neighbor-Joining



S_{12} = sum of branch lengths

$$= L_{1x} + L_{2x} + L_{xy} + \sum_{k=3}^{n} L_{yk}$$

$$= \frac{1}{2} D_{12} + \frac{1}{2(n-2)} \sum_{i=3}^{n} \left(D_{1i} + D_{2i} \right) + \frac{1}{n-2} \sum_{3<i<j} D_{ij}$$

~~$\Sigma (L_{1x} - D_{1x})$~~

$$\min \left[(L_{12} - D_{12})^2 + (L_{13} - D_{13})^2 + (L_{14} - D_{14})^2 , \ldots \right]$$

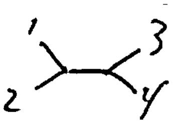
$$\Sigma \hat{L}_{1x} = \Sigma L S = S_{12}$$

Calculates branch lengths as with Fitch-Margoliash

Claim

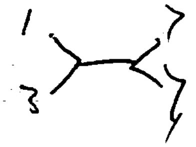
~~NJ if tree is additive~~

NJ --- if distances are additive it will find correct topology.

✓ (A) 

$S_{12} = \frac{1}{2}D_{12} + \frac{1}{4}(D_{13}+D_{14}+D_{23}+D_{24}) + \frac{1}{2}D_{34}$

$= S_{34}$

$S_{13}=S_{24} = \frac{1}{2}D_{13} + \frac{1}{2}D_{24} + \frac{1}{4}(D_{12}+D_{14}+D_{23}+D_{34})$ 

$$\boxed{\begin{aligned} S_{13}-S_{12} &= -1/4\, D_{12} + \tfrac{1}{4}D_{13} + \tfrac{1}{4}D_{24} - \tfrac{1}{4}D_{34} \\ &= \tfrac{1}{4}(D_{13}+D_{24}-D_{12}-D_{34}) \end{aligned}}$$

$S_{13}-S_{12} > 0 \Rightarrow$ tree A is correct

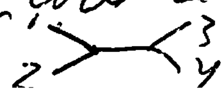
$D_{12}+D_{34} < D_{13}+D_{24}$ similar to before for additive tree

$D_{12}+D_{34} \le D_{13}+D_{24} = D_{14}+D_{23}$

∴ for $N=4$ NJ gives correct tree

w/ non-additive

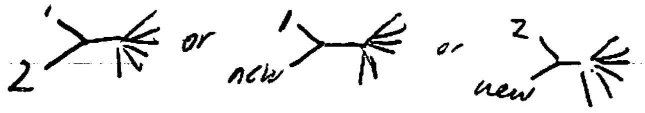
$D_{12}+D_{34} \le D_{13}+D_{24} \le D_{14}+D_{23}$ } one will be true

$D_{12}+D_{34} \le D_{14}+D_{23} \le D_{13}+D_{24}$ } if 

2/16

NJ
11 species

- induction from $\underline{n-1}$ to n
- 3 possibilities



- if S_{12} is the smallest

consider $S_{1n} - S_{12} = \frac{1}{2} D_{1n} + \frac{1}{2(n-2)}\Big[(D_{12}+D_{13}+\dots D_{1n-1}) + (D_{2n}+D_{3n}+\dots D_{3n-1})\Big]$

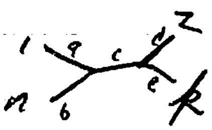
$$+\frac{1}{n-2}\Big[(D_{23}+D_{24}\dots D_{3n-1}) + (D_{34}+\dots D_{3,n-1}) + (\dots\Big]$$

$$-\frac{1}{2} D_{12} - \frac{1}{2(n-2)}\Big[(D_{13}+D_{14}\dots D_{1n}) + (D_{23}\dots D_{2n})\Big]$$

$$+\frac{1}{n-2}\Big(D_{34}+D_{35}+\dots D_{3n}\dots + D_{45}+\dots D_{4n} \dots\Big)$$

$$S_{1n} - S_{12} \boxed{= \frac{1}{2n-2}\sum_{k=3}^{n-1}(D_{1n}+D_{2k}-D_{12}-D_{kn})}$$

~~$S_{1n} - S_{12}$~~

Assume 

$$D_{1n} + D_{2k} - D_{12} - D_{kn} = a+b+d+e-(a+c+d)-(b+c+e)$$
$$= -2c$$

$\therefore \sum_{k=3}^{n-1} -2c$ is always < 0

$\therefore S_{1n} - S_{12}$ would be > 0

2/16/95

<u>Fitch</u>, Doolittle & Feng (JME 18:30. 1981)

<u>Neigborliness</u>

- give two neighbors a <u>+1</u> score for <u>each</u> set of four <u>sequences</u> with that pair in which those <u>two</u> are closest

~~[illegible]~~ $\binom{n-2}{2}$ # of choices

- use this matrix of neighborliness for tree making

Parsimony · Stewart 1993 Nature 361:603

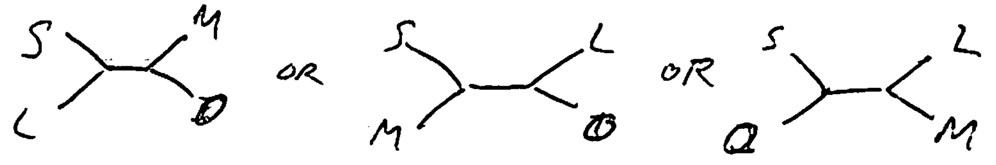
① Given an alignment ... each aligned position is represented by a tree
② Reconstruct internal nodes of tree & get fewest substitutions
③ Add up over all columns

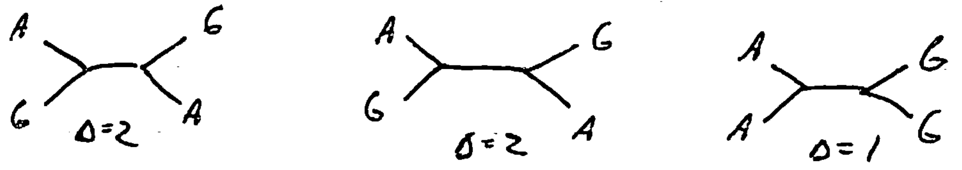
Predict True Tree = tree w/ minimal substitutions

But ... must look at all trees

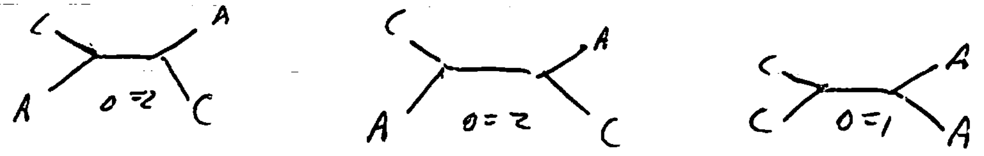
EXAMPLE

	Non Inf		Non Inform.				Non Infm			Non infom
shark	G	A	T	C	C	T	A	G	G	C
lungfish	G	G	T	C	A	C	A	T	G	T
monkey	G	G	T	C	A	T	A	T	C	T
outgroup	G	A	T	A	C	C	A	G	C	A
	1	2	3	4	5	6	7	8	9	10

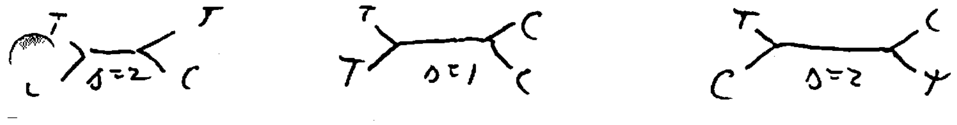




~~P~~ Position 2



position 5



position 6

Σ=9 Σ=9 Σ=7

Maximum likelihood

phylogenetic __or__ phenetic trees

__Principle...__

Observed data - D

- several alternative probability models (e.g. ... diff. trees) M_i
- $P(D|M_i)$ = prob. of observing D under model M_i
- $P(M_i|D)$ = likelihood of $\underline{M_i}$ given

$$= \frac{P(M_i, D)}{P(D)} = p\,\frac{P(D|M_i)\,P(M_i)}{\sum_i P(D|M_i)\,P(M_i)} = \text{BAYES FORMULA}$$

$P(M_i)$ = a priori probabilities

= ∴ assume all $P(M_i)$ equally likely

$$= \frac{P(D|M_i)}{\sum P(D|M_i)}$$

∴ Most likely model is the one that maximizes $P(D|M_i)$

__Example__

- coin tossing
  - if prob(H) = p = M_p
  - if n tosses w/ k heads = D

$$L = Prob\,(D|M_p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\Downarrow$$

$$\log L = \log\binom{n}{k} + \log p^k + \log(1-p)^{n-k}$$

use __log__ Likelihood

WHEN $= \frac{d \log L}{dp} = 0 = \cdots$

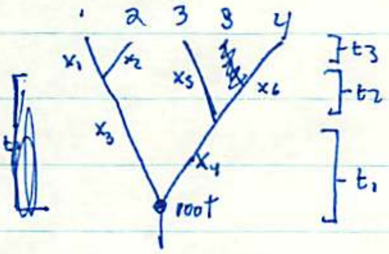
$$\Downarrow$$

$$\hat{p} = \frac{k}{n}$$

MAX. LIKELIHOOD IN PHYLOGENY

① GIVEN TREE ··· ESTIMATE BRANCH LENGTHS

(LANGLEY & FITCH JME 3:161)



- x_i = # of substitutions
- times = unknown

model = poisson process for substitution events

$$P_t(X=x) = \text{prob. of } x \text{ subs. in time } t$$

$$= \frac{\lambda t^x e^{-\lambda t}}{x!}$$

ASSUMPTIONS

- events in one time period independent of others
- linearity for small times
- λ is constant

ESTIMATION

- can only calculate $\lambda^t = v$

LIKELIHOOD $= \sum L_1 \cdot L_2 \cdot L_3 L_4$ $\quad L_1 = \frac{v_1^{x_4} e^{-v_1}}{x_4!}$ $\quad L_2 = \frac{(v_1+v_2)^{x_3} e^{-(v_1+v_2)}}{x_3!}$

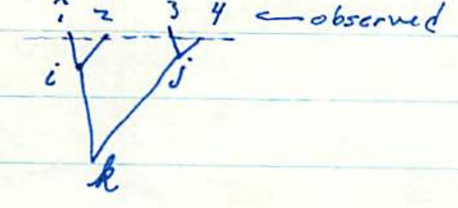
$$L_3 = \frac{v_3^{x_1+x_2} e^{-2v_3}}{x_1!\, x_2!} \qquad L_4 = \frac{(v_2+v_3)^{x_5+x_6} e^{-2(v_2+v_3)}}{x_5!\, x_6!}$$

MAX LIKELIHOOD = take derivative $\frac{d}{dx_1} \frac{d}{dx_2} \frac{d}{dx_3} \frac{d}{dx_4}$
= set to zero
= solve

2) GIVEN SEQS -- ESTIMATE TOPOLOGY & BRANCH LENGTHS

- specify ... $P_{ij}^{\,t}$ = prob. in time t that get $i \to j$ change

- likelihood can be specified for <u>any</u> topology .. on a per nucleotide basis



$$L = \sum_{k=1}^{4}\left(P_k \cdot \left(\sum_{i=1}^{4} P_{ki}^{\,t_i} P_{ix}^{\,t_1} P_{iy}^{\,t_2}\right)\left(\sum_{j=1}^{4} P_{kj}^{\,t_j} P_{jz}^{\,t_3} P_{jv}^{\,t_4}\right)\right)$$

One parameter model $P_{ij}^{\,t} = (1-e^{-\lambda t})P_j \quad i \neq j$

$P_{ii}^{\,t} = e^{-\lambda t} + (1-e^{-\lambda t})P_i$ $\qquad e^{-\lambda t}$ = prob of <u>no</u> event

Algorithm

① sum up log-likelihoods over all sites

② find maximum in terms of $v = \lambda t$

③ get most likely tree

Consistency checks

① bootstrapping ..resample columns w/ replacement

<u>-but...</u>

Pattern freq. distribution

$$P\left(\begin{matrix}\text{pattern } P \\ \text{occurs } x \text{ times}\end{matrix}\right) = ?$$

Definition: sequence S is in state (x, i) if it contains P x times and ends on P_i

$i = 0 \dots\, m-1$

State Δ $f(x)$

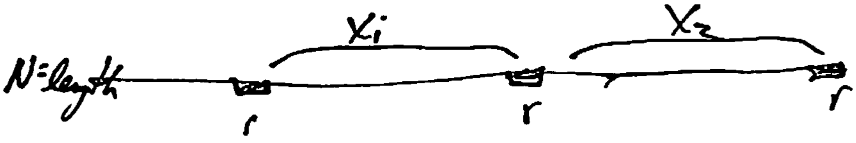
$$s(x, i : j) = \left\{\begin{matrix} x_{\epsilon}, t_{ij} \\ x+1, i_x \end{matrix}\right\} \begin{matrix} t_{ij} \neq m \\ t_{ij} = m \end{matrix}$$

To study words depends on what order of markov model

1 ORDER

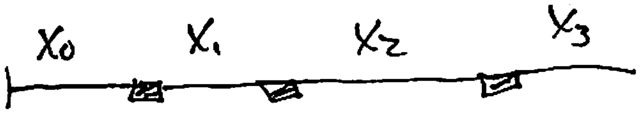
- word = AAAT
- need A
  C
  G
  T
  AA
  AAT

Distribution: betw patters



K = # of events

- $P(N_n < k) = P(X_1 + \cdots X_{k-1} > n)$
- if N is large, k large X = normally distributed



X_0 = 1st passag

How get Markov model

Multiple Alignments

1) Laurence & Reilly Proteins 7:41
2) Krogh et al JMB 235:1501
3) Laurence et al Science 263:208

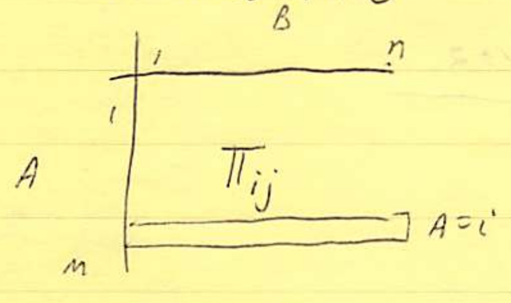
global multiple alignments
- progressive pairwise alignments
- profile methods

local multiple alignments = motifs ??
- e.g - Protein - DNA binding sites
- promoters

Cannot do pairwise alignments because each pw. will come up w/ diff. regions

TWO FUNDAMENDAL CONCEPTS

① conditional probability
- two events A & B



- cond. prob = prob $\{B=j \mid A=i\}$

= ~~prob $\{B=j \cap A=i\}$~~

$$= \frac{\pi_{ij}}{\sum\limits_{j=1 \ldots n} \pi_{ij}}$$

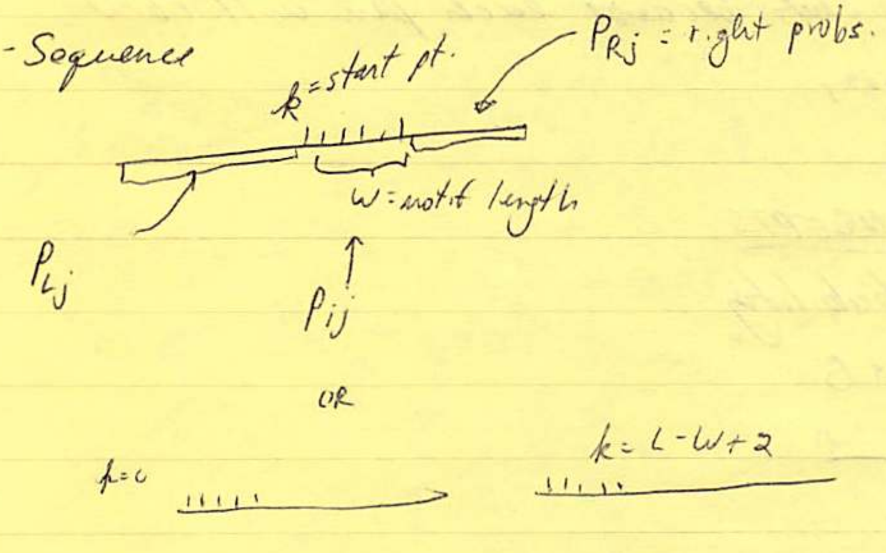
② $f(x_1, x_2 \dots x_n) = \sum_{i=1}^{i=n} p_i \log x_i$ w/ $\sum x_i = 1$

this $f(x)$ has maximum $x_i = \frac{p_i}{\sum_{i=1}^{n} p_i}$

A)-Motif has no gaps
-Some seqs have motif; some don't

n seqs $\{S_1 \cdots S_n\}$
length $\{L_1 \cdots L_n\}$

- Sequence



P_σ = prob. that seq has motif

likelihood
want to maximize $\mathcal{L}_\theta(S)$ $\theta = P_{LJ}, P_{ij}, P_{rj}, P_\sigma$

EM Algorithm

- start w/ θ
- finds θ' such that $g(\theta')$ is $\geq g(\theta)$
- the way to find the next θ'

$$\log g_\theta(S_s) = \log h_\theta(S_s, k) - \log w_\theta(k/S_s)$$

$$\log g_{\theta'}(S_s) - \log g_\theta(S_s) = \frac{\cancel{\log} h_{\theta'}(S_s, k)}{h_\theta(S_s, k)} - \log \frac{w_{\theta'}(k/S)}{w_\theta(k/S)}$$

- multiply by $w_\theta(k/S_s)$ & sum ($=1$ because $\Sigma=1$)

$$\log g_{\theta'}(S_s) - \log g_\theta(S_s) = \sum_k w_\theta(k/S_s) \log \frac{h_{\theta'}(S_s, k)}{h_\theta(S_s, k)}$$

$$- \sum_k w_\theta(k/S_s) \log \frac{w_{\theta'}(k/S)}{w_\theta(k/S)} \quad > 0 \text{ by eq}^s\ 1$$

$g_{\theta'}(S_s) \geq g_\theta(S_s)$

if θ' such that max $\Sigma\, w_\theta(k/S_s) \log h_{\theta'}(S_s, k)$

max from eqn 1

$$= \sum_j \sum_k w_\theta(k/S_s)\, n(L_s k_j) \log P'_{L_j}$$

$$+ (\text{fr right}) + \Sigma\, w_\theta(k/S_s) \log(w_{S_{\theta'}}(k)$$

max when $P'_{LJ} = \dfrac{\sum_k w_\theta(k/S_s)\, n(L_s k_J)}{\sum_J \sum_k w_\theta(k/S_s)\, n(L_s k_J)}$

Algorithm for searching for max θ

$$h(S_s/K) = \prod_j P_{LJ}^{\overset{\text{\# of times J in left}}{n(Lskj)}} P_{RJ}^{(nRskj)} \prod_j P_{ij}^{\overset{\text{is }\underline{J}\text{ at position}}{(nskij)}}$$

$$h_\theta(S_s,k) = h_\theta(S_s/K)\cdot(\text{prob motif is at } K)$$

$$= h_\theta(S_s/k)\,\underline{w_{s\theta}(k)}$$

$$w_{s\theta}(k) = P_J\left(\frac{1}{L_s - w + 1}\right) : k = 1, 2 \ldots$$

$$\frac{1-P_J}{2} : k = 0 \;\underline{\text{OR}}\; k = w - L + 2$$

$$\boxed{g_\theta(S_J) = \sum_k h_\theta(S_J, k)} - \text{MAXIMIZE}$$

$$w_\theta(k/S) = \frac{h_\theta(S_J,k)}{g_\sigma(S_J)}$$

to do this.. start w/ θ, search for local maxima

repeat

EM Continued

-for S sequences replace $\sum_k$ w/ $\sum_s \sum_k$

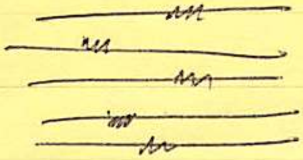
-interpretation

for any parameters...

-can calc. prob. that motif is at k / $w_\theta(k|S_s)$

-can then calc. new parameters from observed freqs. of $L_j R$...

Stochastic Analog = GIBBS Sampling



try to maximize signal vs. background score

signal = P_{ij}

background = P_J

- for position k $A_k = \dfrac{\prod P_{ij} = \text{signal score}}{\prod P_j = \text{background score}}$

∴ want to maximize $\sum_s \sum_i \sum_j n\, k_s\, ij \log \frac{P_{ij}}{P_j}$ over all k_s

① choose random start

② leave 1 sequence out

③ scan that sequence for new A_k using weights from other seqs

④ new parameters

Doug Brutlag Correlations → Structure in Biological Sequences

Molecular Biology is a Information Science

① DNA → RNA → PROTEIN → FUNCTION

② Gentic info → molecular structure → biochemical function → Biological Behavior

Problems

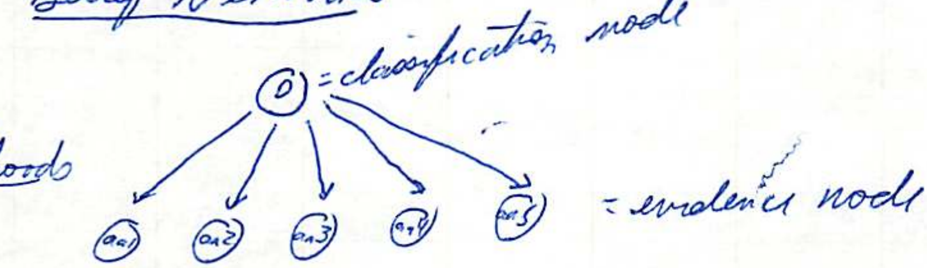
- genetic info is redundant
- structural info is redundant
- multiple featuures encoded by 1 sequence
  - protein sequence
  - folding
  - tl & tx rate

Representation

- most representations assume that sites are independent

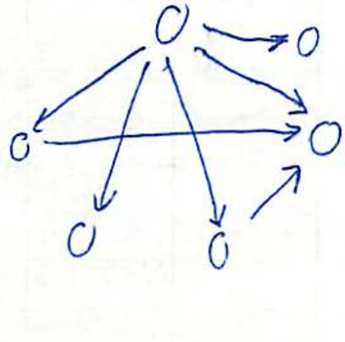
Belief Networks

- see Neapolitan



see Protein Science
send email
to brutlag@cmgm

can add correlations



α-helix

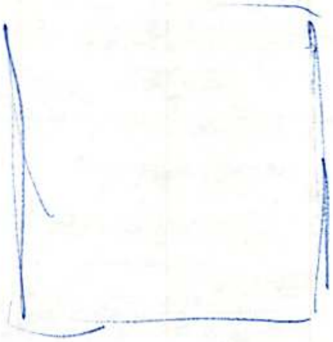
- in 3D space residue i is closest to i+3, i+4
· took a reduced set w/ ~~[illegible]~~ no homologs

Test of correlations

① chi-squared

② mutual information

③ monte-carlo simulations



Examples

i, i+4	D	not D
K		
not K		

overrepresented
KD EK SA
KE FM GA
LL IL PF

underreps
KL

- removed these helices
- appears these aa's interact
③ RANDOMIZED BONDS ...

Generalized

-reduced alphabet size
by classifying into different alphabets

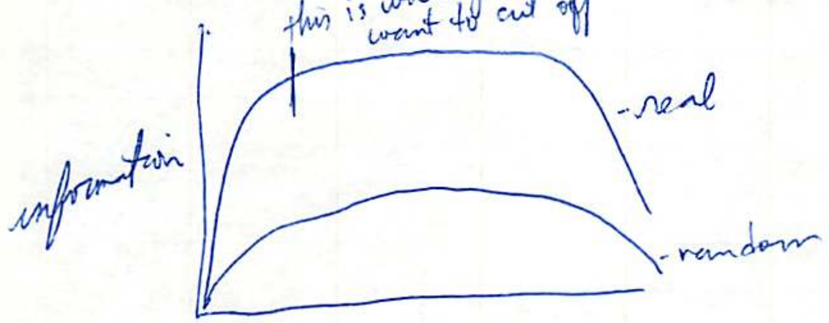
-convert aa into # (parametric) & look
for correlation coefficients

① Do w/ 20×20 alphabet

repeat -pick most $\binom{20}{2}$ similar

- group them

(but what about forcing groups)


this is where you want to cut off
information
-real
-random