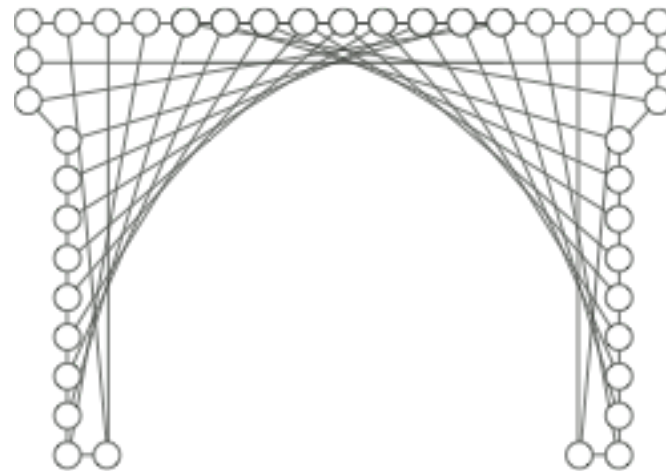


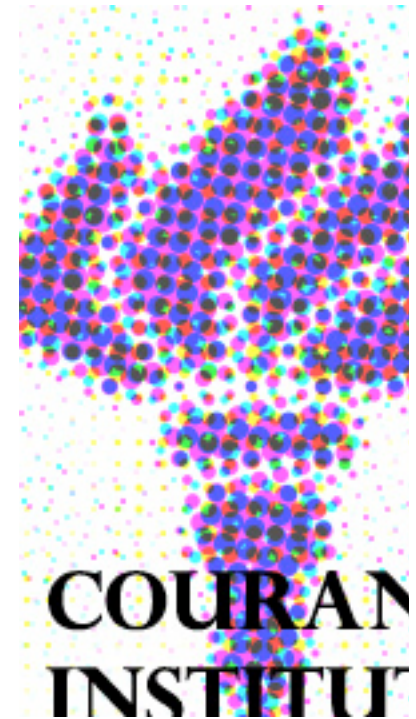
Bioinformatics

Richard Bonneau

Lecture 2: sequence alignment I.



NEW YORK UNIVERSITY
CENTER FOR COMPARATIVE
FUNCTIONAL GENOMICS



COURAN
INSTITUTE

Associated reading.

- Durbin, Eddy, et al. Ch. 2
- Papers for discussion section:
 - Gapped Blast +PSI-BLAST Paper: [Nucleic Acids Res.](#) 1997 Sep 1;25(17):3389-402
 - [Proc Natl Acad Sci U S A.](#) 1993 Jun 15;90(12):5873-7.

Pairwise Alignments I

- What do you need: i. definition of alignment ii. score to rank alignments iii. algorithm to search for alignments iv. score significance of match.
- This lecture
 1. additive scoring scheme, substitution matrices (ii above)
 2. alignment algorithms
 3. judging match, calculating significance
 4. heuristic tools, BLAST, gapped-BLAST, FASTA, PSI-BLAST
 5. Useful links

- Bayes Rule:

$$P(y | x) = \frac{P(x | y)P(y)}{P(x)}$$

- Small amount of probability:

$$P(x | \vec{Y}) = P(x | y_1)P(y_1) + \dots + P(x | y_n)P(y_n)$$

- Estimating P given 1 sample and a distribution, example with normal distribution, t-test. We'll use different distribution today.

(i.) alignment def

- sequence i : ABAACABA
- sequence j : ABAAABA
- alignment :
abaa-caba
| | | | . . | | |
abaaa-aba

(ii.) ranking alignments

alignment : abaa-caba
 | | | | . . | | |
 abaaa-aba

One simple score:

alignment to gap = mismatch

$$S(x_i, y_j) = \left\{ \begin{array}{l} 2 \text{ if match} \\ -1 \text{ mismatch} \end{array} \right\}$$

Sum over **S**

- Better score for matches
approximates $p(x,y|\text{Match})$:

$$P(x,y|\text{Random}) = \prod_i q_{x_i} \prod_j q_{y_j}$$

$$P(x,y|\text{Match}) = \prod_i p_{x_i y_i}$$

$$\frac{P(x,y|M)}{P(x,y|R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_j q_{y_j}}$$

$$\log \left(\frac{P(x,y|M)}{P(x,y|R)} \right) = S(x,y)$$

```

FRHLVLLDALLKRH.SVSAAAR.ELDL...POPTASHGLARI
PRLLRFLFDALFTTG.SVTKAAE.RLCQ...SQPTVSIWLARI
PKLLQLFDVLYQCR.SVTRAAE.QLGQ...SQPTISIWLARI
FMHLTVFQALMRTK.SVSLAAE.MLDM...POPTLSRHLKQI
M...VFDALYRHG.SACKAAH.ALSM...POPTLSRWLAQI
LNLLVSLDILLAEN.SVSRAAE.RLCL...SNSAMSRILSRI
LNLLFALDVLLEAG.SVARAAR.RLRL...SPSAMSRTLARI
LNLLVTLNALLAEG.SVAGAAR.RLGL...SPSAMSRLARI
MNLVALDALLDEG.SVVGAAQ.RMNL...SPAAMSRTLGRI
LNMLVALNALIEER.SVTGATR.RLHL...SSPVMSRPLARI
LNLLAALDALLEEG.SVAGAAA.RLHV...TAPAMSRSLGRI
LNLLVCLKVLIIEEL.SVTRAAA.RLCL...SQSAVSKSLAKI
LNLLVCLKVLFEEEL.NVTRTAH.RLCL...SQSAVSKALAKI
LNLLVILKVLLLEEQ.SVTRAAE.RLHI...QSALSLSLNRI
LNLLKVLAVMLEER.SVARCAE.RLFV...SPSAVSHALAKI
LNLLVALRVLIEES.SVSKAAV.RLNL...SQSAMSRLVGR
LNLLKALDALLDER.NVTRAAV.RLAI...TQPAMSGMLTRI
LNLLMIFDAILAEG.HVTRAAE.RLAM...SQSAVSKGLAQI
LNLLPILLALHDAR.SVSMAAQ.QLGM...SQPCGVSTALAKI
LNLLVTLDVLLTEH.NVTRAAE.RLNF...SQPSVSVHLAKI
LNLLVSLDVLLEEC.NVTRAAQ.RLHV...SQPALSAQLGRI
LNLLVTLAVLLRER.SVSRAAA.CLHL...GQPAVSGALGRI
LNLLVVLDALLSER.HVSRAAQ.RLAM...SQPAVSHALGRI
LNLLRVLDALLRER.NVSRAAE.RLAL...SQPAVSNALNRI
LNLLRVFDMLREQ.NVSRAAE.RLAL...TOPTVSNALARI
LNLLTALDALLREA.NVSRAAM.RIGL...SQPATSHALQRI
LNLLVALDILLEEQ.NITRAAE.RLHM...TQSATSQVLRGRI
LNLLVVLDALLDEA.HVSRAAD.RLGL...SQPAASAALQRI
LNLLVVLNALLDER.SVTRAAE.RLGM...SQPAVSRALARI
LNLLVTLQALMTEK.HISRTAM.RLHK...SQPAISHALARI
LNLLLALHALLSER.HVTRAAE.RLHR...SQPAVSHALAQI
IKLLRLLVLMSEK.SVSRTAD.RLDV...SQPAVSHALARI
LNLLRSLVLLLEEC.HVSRAAD.RLHI...TQSAMSRLAQI
LNLLRSLHVLLEEC.HVSRTAE.RLHV...TQSAVSRQLAQI
LNLLRSLLEALLETR.NLTRAAA.RLGL...TQSAMSRLVQI
LAHLRTL D HLLQLK.NLSHAAE.RLCV...SQSALSRLQAH
LDLLVAADVILECE.NLTLAGE.RLGL...SQPAVSRTLGRI
LNLVVALRALLEER.NVTRAGQ.RVGL...SQPAMSAALARI
LNLLVALEALLEYR.NVTHAGQ.HIGR...SQPAMSRALGRI
LNLLVDLEALLQYR.HITQAAQ.HVGR...SQPAMSRALSRI
LNTLLALEALLEHR.NVTQAAE.HLGL...SQPSVSRALIRI
LNLLTILEKLLIHK.HISQAAQ.ALNM...SQPAVSRALMRI
LNTLLVVDALLAER.NLSAAAR.RINT...SQPAMSAAVARI

```

Better score for matches approximates $p(x,y|Match)$:

Where do we get the
“|M” in $P(x,y|M)$!!

BLOSUM 50,62:
From high quality
alignments
with no redundancy
> 50, 62% identity

PAM, others

```
FRHLVLLDALLKRH.SVSAAAR.ELDL...PQPTASHGLARLRK.ALGDP...L
PRLRLFDALFTTG.SVTKAAE.RLGG...SQPTVSIWLARLRQ.ELDDP...L
PKLLQLFDVLYQCR.SVTRAAE.QLGQ...SQPTISIWLARLRE.QLNDP...L
FMHLTVFQALMRTK.SVSLAAE.MLDM...PQPTLSRHLKQLRE.HFGNQ...L
M...VFDALYRHC.SAGKAAH.ALSM...PQPTLSRWLAQLRT.HFDDP...L
LNLLVSLDTLLAEN.SVSRAAE.RLGL...SNSAMSRILSRIRE.IFGDP...L
LNLLFALDVLLAEG.SVARAAR.RLRL...SPSAMSRTLARLRE.ATGDP...L
LNLLVTLNALLAEG.SVAGAAE.RLGL...SPSAMSRLARLRA.TMNDP...L
MNLLVALDALLDEG.SVVGAAQ.RMNL...SPAAMSRTLGRIRE.ALGDP...L
LNMLVALNALIEER.SVTGATR.RLHL...SSPVMRPLARLRE.TLDDP...L
LNLLAALDALLEEG.SVAGAAA.RLHV...TAPAMSRSLGRIRR.TTGDQ...L
LNLLVCLKVLIIEEL.SVTRAAA.RLCL...SQSAVSKSLAKLRE.QFDDP...L
LNLLVCLKVLFEEEL.NVTRTAH.RLCL...SQSAVSKALAKLRQ.QFDDP...L
LNLLVILKVLLEEQ.SVTRAAE.RLHI...SQSALSLSLNRRLD.TLDDP...L
LNLLKVLAVMLEER.SVARCAE.RLFV...SPSAVSHALAKLRQ.MFSDP...L
LNLLVALGRVLIIEES.SVSKAAV.RLNL...SQSAMSRLGRIRD.LFGDP...L
LNLLKALDALLDER.NVTRAAV.RLAI...TQPAMSGMLTRLRD.SFDDP...L
LNLLMIFDAILAEG.HVTRAAE.RLAM...SQSAVSKGLAQLRQ.AFGDP...L
LNLLPILLALHDAR.SVSMAAQ.QLGM...SQPCVSTALAKLRT.AFGDP...L
LNLLVTLQALMTEK.HISRTAM.RLHK...SQPAISHALHLRD.IFNDP...L
LNLLVSLDVLLIEEC.NVTRAAQ.RLHV...SQPALSAQLGRLRH.LFDDP...L
LNLLVTLAVLLRER.SVSRAAA.CLHL...GQPAVSGALGRLRE.LFDDE...L
LNLLVVDALLSER.HVSRAAQ.RLAM...SQPAVSHALGRLRE.LFGDP...L
LNLLRVFDMLLREQ.NVSRAAE.RLAL...TQPVTSNALNRIRD.QLGDP...L
LNLLTALDALLREA.NVSRAAM.RLGL...SQPATSHALQRLRD.IFGDP...L
LNLLVALDILLEEQ.NITRAAE.RLHM...TQSATSQVLGRLRT.FFEDE...L
LNLLVVDALLDEA.HVSRAAD.RLGL...SQPAASAALQRCRH.LFRDE...L
LNLLVVLNALLDER.SVTRAAE.RLGM...SQPAVSRALARLRA.LFSDA...L
LNLLVTLQALMTEK.HISRTAM.RLHK...SQPAISHALHLRD.IFNDP...L
LNLLLALHALLSER.HVTRAAE.RLHR...SQPAVSHALGRLRE.LFGDP...L
IKLLRITLLVLMSEER.SVSRTAD.RLDV...SQPAVSHALARLRI.LFDDP...L
LNLLRSLVLLIEEC.HVSRAAD.RLHI...TQSAMSRLAQLRE.LCSDP...L
LNLLRSLHVVLLIEEC.HVSRTAE.RLHV...TQSAVSRQLAQLRE.LCGDP...L
LNLLRSLAALLETR.NLRSAAE.RLGL...TQSAMSRHLVQLRA.QFQDP...L
LAHLRTLDHLLQK.NLSHAAE.RLGV...SQSALSRLAHLRE.AFDDP...L
LDLLVAADVILECE.NLTLAGE.RLGL...SQPAVSRTLGRLRE.AFDDP...L
LNLLVVALRALLEER.NVTRAGQ.RVCL...SQPAMSAALARLRR.HFDDD...L
LNLLVALEALLEYR.NVTHAGQ.HIGR...SQPAMSRALGRLRG.LFNDD...L
LNLLVDLEALLQYR.HITQAAQ.HVGR...SQPAMSRALSRLRG.MLKDD...L
LNTLLALEALLEHR.NVTQAAE.HLGL...SQPSVSRALIRLRE.VFNDD...L
LNLLTILEKLLIHK.HISQAAQ.ALNM...SQPAVSRALMLRRE.QFGDP...L
LNLLVVDALLAER.NLSAAAR.RINL...SQPAMSAAVARLRE.FFGDE...L
LNLLVVDALLTER.TLTAAAS.SINL...SQPAMSAAVARLRE.YFNDE...L
LNLLVAFDAVMTER.SVTAAAR.SINL...SQPAMSAAIARLRL.YFGDE...L
```


- Better score for matches
approximates $p(x,y|\text{Match})$:
- Scoring gaps, gap penalties:
- g - length of gap; d - gap open
 e - gap extend

$$\gamma(g) = -eg \implies \textit{linear}$$

$$\gamma(g) = -e(g - 1) - d \implies \textit{afine}$$

(iii.) Algorithm for finding optimal matches

Cost of trying every possible alignment exhaustively

$$\text{if } : n_i = n_j = n$$

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

Brute force wont cut it...

(iii.) Algorithm for finding optimal matches

Dynamic Programming:

-global = Needleman-Wunsch 1970

Gotoh 1982

-local = Smith Waterman

(iii.) Algorithm for finding optimal matches
(global)

We base larger global alignments on smaller preceding alignments.

Best alignment for x - up to $x_i \Rightarrow y$ - up to y_j
based on alignment up to x_{i-1}, y_{j-1}

(iii.) Algorithm for finding optimal matches
(global)

I. Set $F(0,0) = 0$

	0	c	a	d	b	d	
0	0						
a							
c							
b							
c							
d							
b							

(iii.) Algorithm for finding optimal matches (global)

$F(i,j)$:

3 possible outcomes:

$$x_i \rightarrow y_j, \quad F(i,j) = F(i-1, j-1) + s(x_i, y_j)$$

$$x_i \rightarrow \text{gap}, \quad F(i,j) = F(i-1, j) - d$$

$$\text{gap} \rightarrow y_j, \quad F(i,j) = F(i, j-1) - d$$

$F(i,j) = \max$ of three above

Simple score for this example:

Mismatch = -1

Match = 2

(iii.) Algorithm for finding optimal matches (global)

1. Set $F(0,0) = 0$

2. $F(i,0) = -id$

3. $F(0,j) = -jd$

	0	c	a	d	b	d	
0	0	-1	-2	-3	-4	-5	
a	-1						
c	-2						
b	-3						
c	-4						
d	-5						
b	-6						

(iii.) Algorithm for finding optimal matches (global)

1. Set $F(0,0) = 0$

2. $F(i,0) = -id$

3. $F(0,j) = -jd$

4. Fill in $F(i,j)$ from
top-left to
bottom-right
(remember
pointers)

	0	c	a	d	b	d	
0	0	-1	-2	-3	-4	-5	
a	-1	-1	1				
c	-2	1					
b	-3						
c	-4						
d	-5						
b	-6						

(iii.) Algorithm for finding optimal matches (global)

1. Set $F(0,0) = 0$
2. $F(i,0) = -id$
3. $F(0,j) = -jd$
4. Fill in $F(i,j)$ from top-left to bottom-right (remember pointers)
5. Traverse to read **alignmnt**

c a d b - d -
 . | . | . | .
 - a c b c d b

Score = 2

	0	c	a	d	b	d	
0	0	-1	-2	-3	-4	-5	
a	-1	-1	1	0	-1	-2	
c	-2	1	0	0	-1	-2	
b	-3	0	0	-1	2	1	
c	-4	-1	-1	-1	1	1	
d	-5	-2	-2	1	0	3	
b	-6	-3	-2	0	3	2	

(iii.) Algorithm for finding optimal matches (global)

$F(i,j)$:

3 possible outcomes:

$$x_i \rightarrow y_j, \quad F(i,j) = F(i-1, j-1) + s(x_i, y_j)$$

$$x_i \rightarrow \text{gap}, \quad F(i,j) = F(i-1, j) - d$$

$$\text{gap} \rightarrow y_j, \quad F(i,j) = F(i, j-1) - d$$

No outcome > 0 , $F(i,j) = 0$

$F(i,j) = \max$ of three above OR 0

0 effectively starts new alignment

Both these algorithms work in $O(nm)$ time
and $O(nm)$ space

Should i explain big O?

(iii.) Algorithm for finding optimal matches (local - smith-waterman)

$F(i,j)$:

3 possible outcomes:

$$x_i \rightarrow y_i, \quad F(i,j) = F(i-1, j-1) + s(x_i, y_j)$$

$$x_i \rightarrow \text{gap}, \quad F(i,j) = F(i-1, j) - d$$

$$\text{gap} \rightarrow y_i, \quad F(i,j) = F(i, j-1) - d$$

No outcome > 0, $F(i,j) = 0$

$F(i,j) = \max$ of three above OR 0

0 effectively starts new alignment

(iv.) calculating/estimating alignment significance (bayes)

We have: $P(x, y | M)$

We want: $P(M | x, y)$

I take it back, i'm skipping some steps here ...
we don't want to be here all day

Bayes rule: $P(M | x, y) = \frac{P(x, y | M)P(M)}{P(x, y)}$

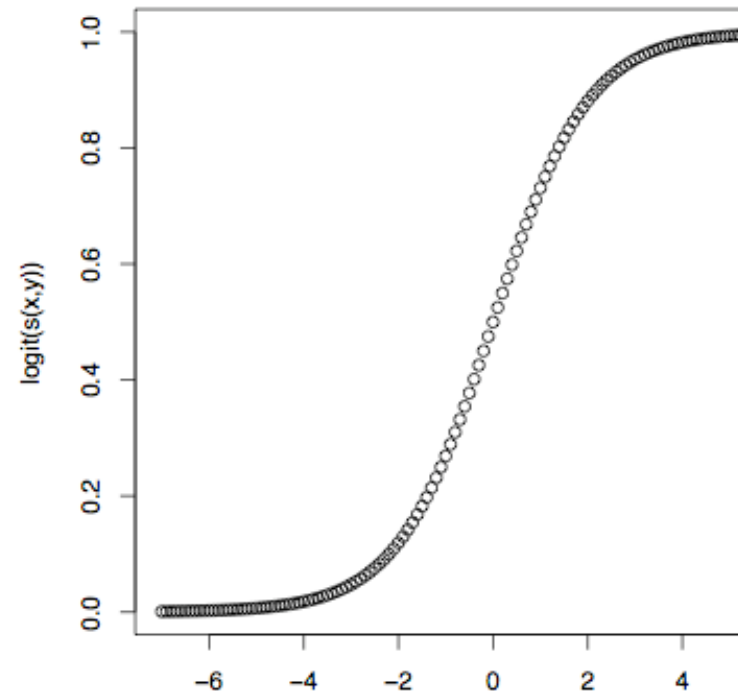
$$P(x, y) = P(x, y | R)P(R) + P(x, y | M)P(M)$$

$$P(M) = \text{prior}$$

$$P(R) = 1 - P(M)$$

$$S' = \log\left(\frac{P(M)}{P(R)}\right) + \log\left(\frac{P(x, y | M)}{P(x, y | R)}\right)$$

$$P(M | x, y) = \sigma(S') = \frac{e^{S'}}{1 + e^{S'}}$$



(iv.) calculating/estimating alignment significance
(extreme value distribution)

$$P(M_N \leq x) \simeq \exp\left(-KNe^{\lambda(x-\mu)}\right)$$

Extreme value distribution

Free parameters:

Ungapped: S Karlin + Altschul analytical

Gapped : - Believed to follow analytical
solution from Karlin (Mott 1992)
- Altschul + Gish 1996 - fit from
large generated dataset.

Next week's reading

- Durbin, Eddy, et al. Ch. 3
- Papers for discussion section: