End of Semester Logistics

711-3: due Sunday, Dec 1, 11:59pm

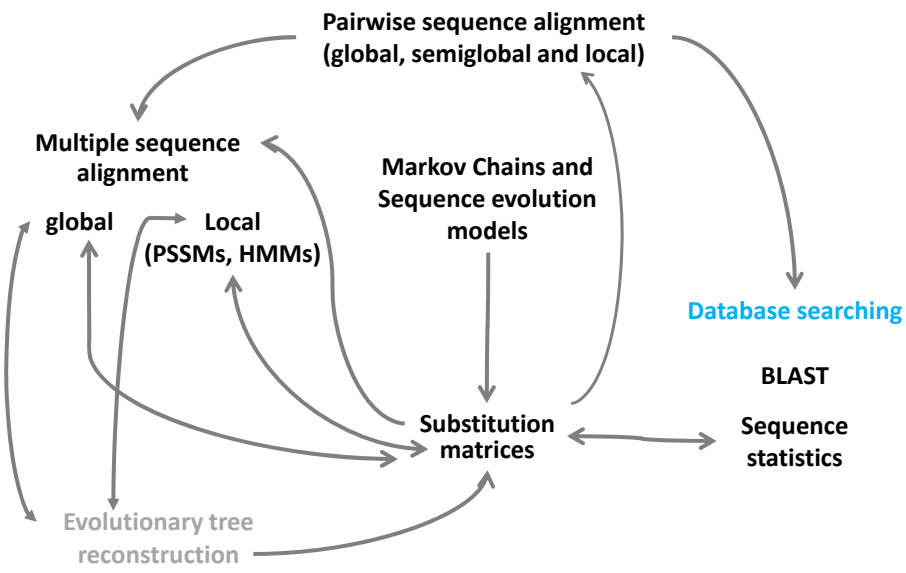
Last day of class: Thursday Dec 5th

PS7/711-4: due Friday, Dec 6, 11:59pm

Review session: Sunday, Dec 8th?

Final exam: Monday, Dec 9th, 1pm-4pm

- Cummulative, emphasis last 3rd of the semester
- Closed book, 2 pages of notes



Searching a sequence database

Input:

- query Q of length *m*
- database D=D1 D2 D3... D_N of length *n*

Search:

for j = 1 to N

- Find best local aligment of Q with Dj
- If "good alignment", add Dj to *Results*

Output: *Results*

PROBLEMS

- Too slow
- What is a "good" alignment?
- Which matrix should you use?
- Which results are trustworthy?
- Can you find all related sequences in the database?

3

- Blast heuristic — Last Thursday
- Blast statistics: How significant is a matching sequence with score *S?* — Today
- How much information is available to distinguish between chance MSPs and MSPs in related sequences? — Next Week
  - Information content of substitution matrices
  - Information content of alignments
- Which substitution matrix will maximize precision and recall?

Terminology

Segment Pair (SP): ungapped local alignment.

Maximal Segment Pair (MSP): an ungapped local alignment that cannot be improved by making it bigger or smaller.

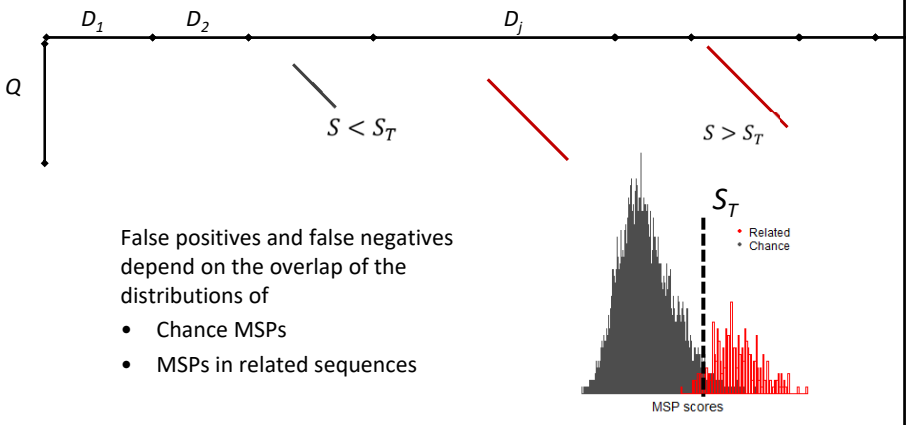
High-Scoring Segment Pair (HSP): an MSP with score at least S_T, where S_T is a user-defined minimum.

Word: String of length *w.* Typically, *w<10.*

Hit: An ungapped alignment of a word in *Q* and a word in *D* with score at least *T.*

6

Given a query sequence, *Q,* of length *m,* and a database sequence, *D*, of length *n,* find all ungapped local alignments with score at least S_T (HSPs)



False positives and false negatives depend on the overlap of the distributions of

- Chance MSPs
- MSPs in related sequences

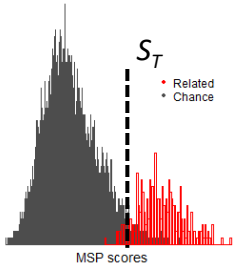
8

Sensitivity and Specificity

False Positives: Unrelated sequences with score ≥ S_T

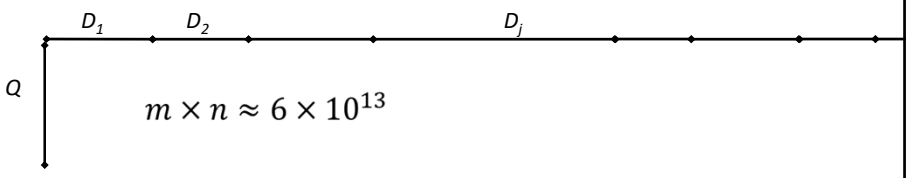
False Negatives: Related sequences with score < S_T

Depend on S_T



Given a query sequence, *Q,* of length *m,* and a database sequence, *D*, of length *n,* find all ungapped local alignments with score at least S_T

For sufficiently large *D,* dynamic programming is too slow.



	Non-redundant (nr) sequence database	
Sequences	Nucleic Acid	Amino Acid
Date:	Nov 18, 2024 4:16 AM	Nov 19, 2024 4:05 AM
Letters:	2,520,486,308,772	321,840,402,579
Sequences:	110,905,049	841,786,231

10

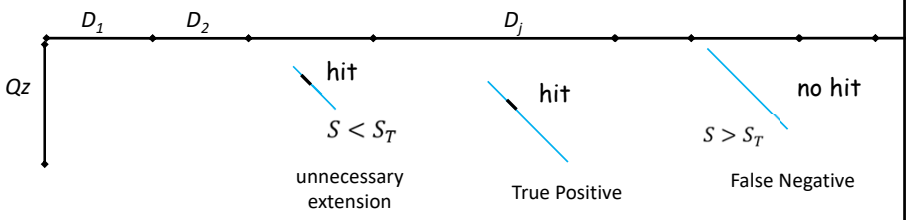
Ungapped BLAST

Basic Local Alignment Search Tool

Altschul *et al*, 90

1. Construct *L:* a list of words of length *w* with score ≥ *T*
2. Scan database *D* for *hits* – instances of words in *L*
3. Extend hits to find MSPs with score > S_T.

Given a query sequence, *Q*, of length *m* and database, *D*, of length *n,* find HSPs



	Non-redundant (nr) sequence database	
Sequences	Nucleic Acid	Amino Acid
Date:	Nov 18, 2024 4:16 AM	Nov 19, 2024 4:05 AM
Letters:	2,520,486,308,772	321,840,402,579
Sequences:	110,905,049	841,786,231

12

Blast90 parameters

False Positives: Unrelated sequences with score ≥ S_T

False Negatives:

1. Related sequences with score < S_T

2. Related sequences that do not contain a hit

Increased running time:
Hits that are not in a high scoring segment pair.

Depend on S_T

Depend on *w* and *T*

Select *w* and *T* to get the best balance between *false negatives* and *efficient running time*.

Problems with Blast 90

1. Construct *L:* a list of words of length *w* with score ≥ *T*
2. Scan database *D* for *hits* – instances of words in *L*
3. Extend hits to find MSPs with score > $\boldsymbol{S_T}$.

90% of runtime

1. Running time: Too many unnecessary extensions
2. Ungapped extensions: related sequences with several short regions of similarity are not retrieved

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

Altschul *et al*, 97

- Gapped BLAST
- Two-Hit BLAST
- PSI-BLAST

2. Ungapped extensions: related sequences with several short regions of similarity are not retrieved

```
43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
   F  L +   V+ +PK+ AH +KV          L + GE V  LD   G+
45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE  90

91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
   +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

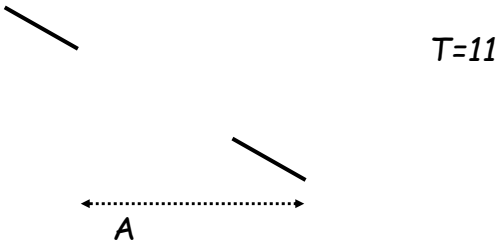
Altschul *et al*, 97

An example: This alignment has two conserved regions connected by gapped region

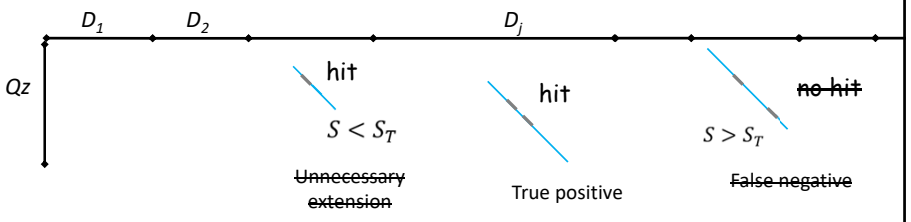
Two-Hit BLAST

Altschul *et al*, 97

- Reduce threshold *T* to obtain *more* hits
- Only trigger an ungapped extension if there are *two hits* on the *same diagonal* within distance ***A***



Two-hit Blast



Reduces

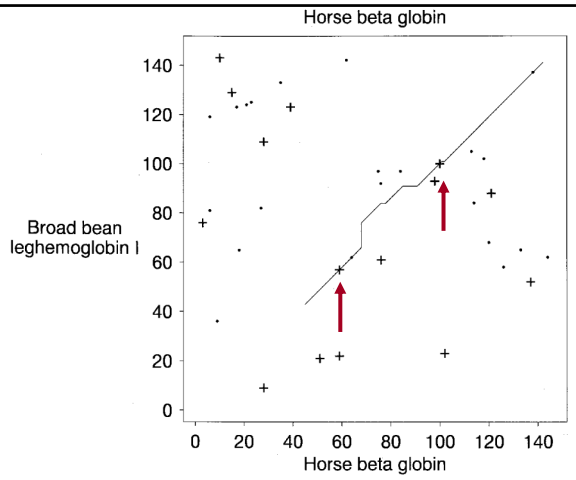
- unnecessary extensions
- false negatives

19

+ 15 hits with score ≥ 13

- 37 hits with score ≥ 11

Only two extensions were performed!




```
43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
   F  L +   V+ +PK+ AH +KV          L + GE V   LD   G+
45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHPLDNLKGTFAALSE  90

91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
   +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

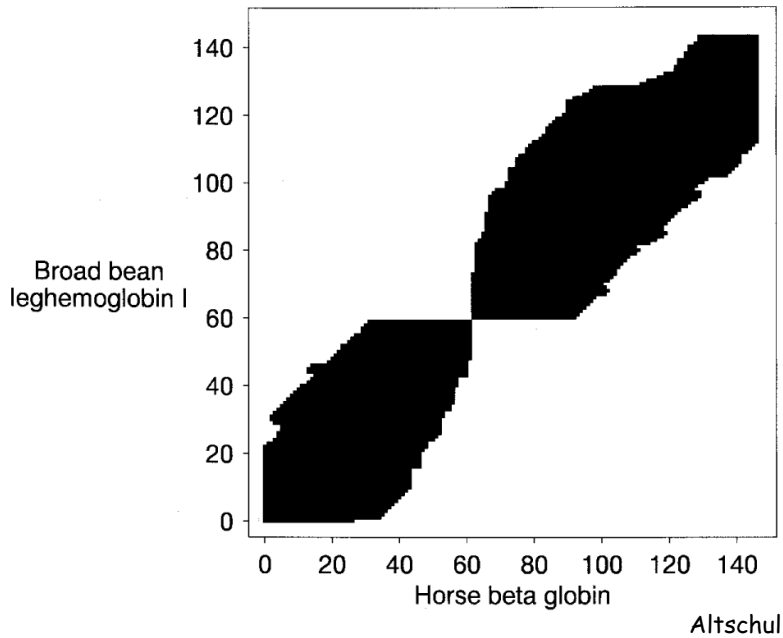
Altschul *et al*, 97

Gapped Extensions

- Find hits of length *w* with similarity threshold *T*.
- If D_j contains
  - *two* hits
  - on *same diagonal*
  - separated by a *distance of at most A*,

  1st test
- Perform an *ungapped* extension using cutoff, *X1*.
- For each ungapped MSP with score S_1, if $S_1 > S_g$, 2nd test
- perform a *gapped extension* with dynamic programming with cutoff *X2*
- If gapped extension score $S_2 > S_T$, report match. 3rd test

cutoff *Xg* limits search space



Altschul *et al*, 97

Two Hit BLAST: Performance

- Reduce threshold *T* to obtain *more* hits
- Only trigger an ungapped extension if there are *two hits* on the *same diagonal* within distance ***A***

- For *w=3, T=11, A=40,*
  - *3.2* times as many hits
  - *0.14* times as many extensions
  - speed up ~ *2X*

Gapped Extensions: Performance

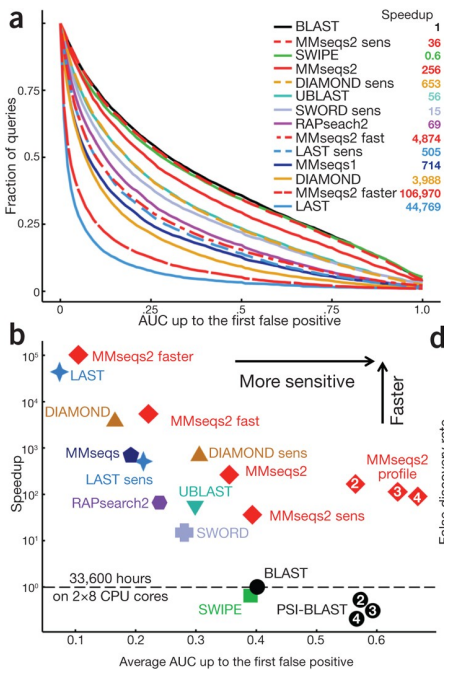
1. Find HSP's, using *ungapped* extensions
2. If HSP score > S_g, perform a *gapped* extension.
3. If gapped extension score $S_2 > S_T$, report match.

- Two Hit strategy reduces the number of ungapped extensions
- Gapped extensions cost 500 times ungapped extensions
- One gapped extension per 4000 ungapped extensions

➢ An additional reduction in running time by more than a factor of 2.

Other DB search methods

- Algorithmic improvements
- often at the expense of sensitivity

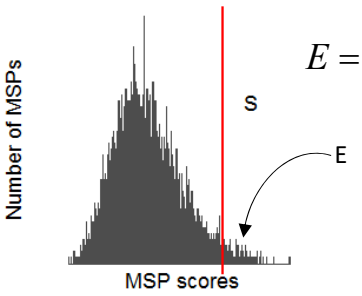


Steinegger, M., Söding, J. *Nat Biotechnol* (2017). https://doi.org/10.1038/nbt.3988

- Blast heuristic — Last Thursday
- Blast statistics: How significant is a matching sequence with score *S?* — Today
- How much information is available to distinguish between chance MSPs and MSPs in related sequences? — Next Week
  - Information content of substitution matrices
  - Information content of alignments
- Which substitution matrix will maximize precision and recall?

BLAST (Karlin-Altschul) Statistics

E = Expected number of matches with score at least *S* under the null model



$$E = Kmne^{-\lambda S}$$

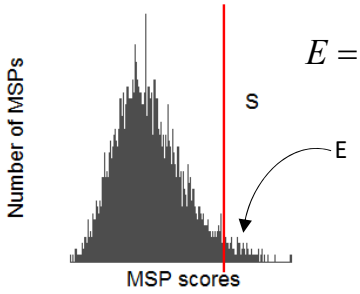
E-values depend on K and λ, which in turn depend on the scoring matrix, *S[i,j]*.

Maximal Segment Pair (MSP): an ungapped local alignment that cannot be improved by making it bigger or smaller.

27

BLAST (Karlin-Altschul) Statistics

E = Expected number of matches with score at least *S* under the null model



$$E = mn2^{-S_b}$$

By normalizing the alignment scores

$$S_b = \frac{\lambda S - \ln K}{\ln 2}$$

we obtain a **"bit score"** S_b and an expression for *E* that is independent of *K* and λ

29

- Blast heuristic — Last Thursday
- Blast statistics: How significant is a matching sequence with score *S*? — Today
- How much information is available to distinguish between chance MSPs and MSPs in related sequences? — Next Week
  - Information content of substitution matrices
  - Information content of alignments
- Which substitution matrix will maximize precision and recall?