

Chapter 2: Sequence Alignment

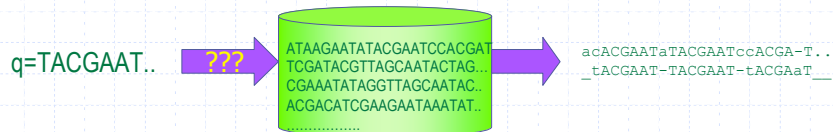
2.3 Searching Sequence Databases; FASTA, BLAST

Prof. Yechiam Yemini (YY)
Computer Science Department
Columbia University

COMS4761-- 2007

The Problem

- How to search a sequence database (DB) for local alignments of a query sequence?
 - E.g., Search a promoter sequence in a DB of 10^5 sequences



- Dynamic Programming is prohibitively complex
- Need techniques that are:
 - Fast: focus search on likely solutions (trade speed for completeness)
 - Tunable: retrieve meaningful alignments (ones with sufficiently high score)
- FASTA & BLAST

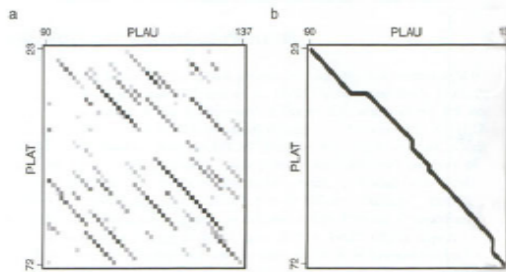
Tutorial: W.R. Pearson "Protein sequence comparison and Protein evolution Tutorial - ISMB2000" (October, 2001)
<http://www.people.virginia.edu/~7Ewrp/papers/ismb2000.pdf>

COMS4761-- 2007

2

Reconsider DP Geometry

- Diagonal matching segments provide the basis for alignments
- Alignment may be viewed as connecting matching diagonals
 - Using mismatched diagonals or horizontal/vertical gapped segments
- Scoring is additive contributions of matching diagonal and connectors
 - Mismatched diagonals or vertical/horizontal gapped segments may reduce the score
 - It is best to focus on high scoring diagonals and use connectors with positive score



COMS4761-- 2007

3

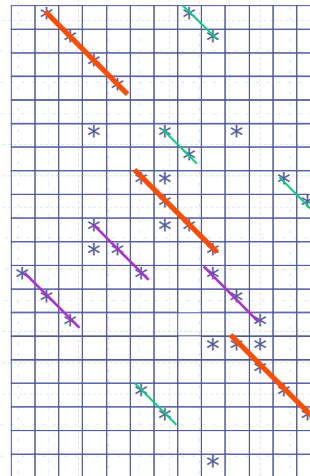
Dot Matrix Heuristics

Rule 1: Find high-scoring diagonals

- Search small diagonal segments
- Extend to max diagonal matches
- Connect diagonals to max score

Rule 2: Focus on meaningful alignments

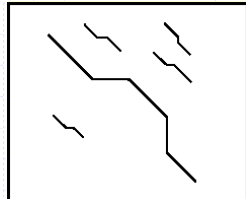
- Filter low-scoring diagonals



COMS4761-- 2007

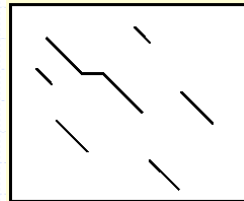
4

Tradeoff: Time vs. Optimality



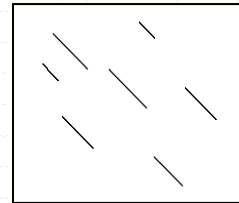
Smith-Waterman

10 min



FASTA

2 min



Blast

20 sec

COMS4761-- 2007

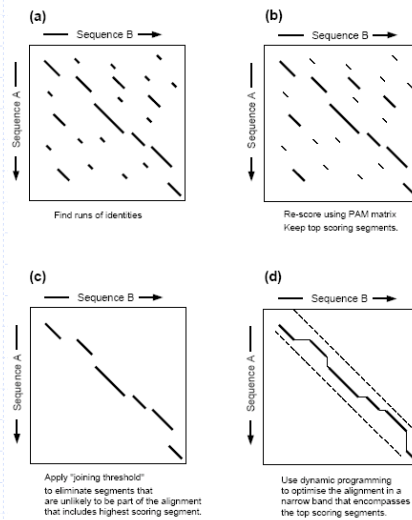
5

FASTA

Key idea (Pearson & Lipman 88):

- Find short diagonals by indexing the DB
- Extend these to high scoring diagonals
- Use DP to connect them

A 4 steps process



COMS4761-- 2007

6

Step (a): Find Diagonal Matches by Indexing

- Key idea: create an index of k-tuples of the DB
 - Scan database to index k-tuples [$k=1..5$]
 - Scan query to index k-tuples
 - Find all diagonal matches of length k by comparing the hash tables
 - Merge these short diagonals into maximal diagonal matches

Example:

Database d: TATCGATCGA
 Position: 1 2 3 4 5 6 7 8 9 10

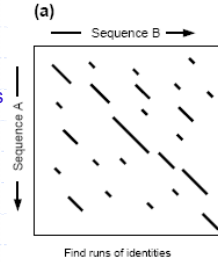
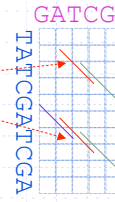
Query q: GATCG
 Position: 1 2 3 4 5

1. Extract index

TAT	1
ATC	2, 6
TCG	3, 7
CGA	4, 8
GAT	5

GATCG	1
TCGAT	2
CGATC	3
GATCG	4
TCGAT	5

0, -4
0, -4
-4



Find runs of identities

2. Find matches

3. Merge diagonal matches

COMS4761--2007

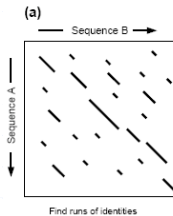
7

FASTA Steps (b-d): Optimize Score

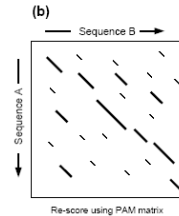
b) Filter low-score diagonals

c) Extend diagonals to max score; keep high-scoring segments

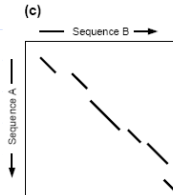
d) Use DP for a narrow band around the high scoring segments



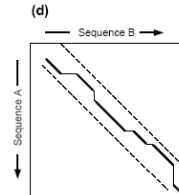
Find runs of identities



Re-score using PAM matrix. Keep top scoring segments.



Apply "joining threshold" to eliminate segments that are unlikely to be part of the alignment that includes highest scoring segment.



Use dynamic programming to optimise the alignment in a narrow band that encompasses the top scoring segments.

COMS4761--2007

8

Example

YOUR EMAIL	SEARCH TITLE	RESULTS	PROGRAM	DATABASES					
<input type="text"/>	Sequence	interactive	fast3 fast3 fasty3	Protein UniProt UniRef100 UniParc					
GAP PENALTIES		SCORES & ALIGNMENTS	KTUP/HISTOGRAM	DNA STRAND					
OPEN	-12	SCORES	10	KTUP	2	DNA STRAND	none	MATRIX	BLOSUM50
RESIDUE	-2	ALIGN	50	HIST	no				
EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE	SEQUENCE RANGE	DATABASE RANGE	MOLECULE TYPE					
10.0	default	START-END	START-END	Protein					

Enter or Paste a Sequence in any format.

```
>Mus musculus protein sequence  
msaikkhmqe lkidkenvid raeqaeqk qeerskyle  
delatmqkkl kgtedeldky sealkdaqk lelaekkaad  
aeaevaslnr riqiveeid eaqratal qlieeakaa  
deaeqgkvi enalkdeek meliqike  
akhiaeead kyeevarkiv sieqierete eraeiaekc  
seleeeiknv tnikleaq nekysqekd yeeikilt  
kikeaetrae faersvakie ktaddiedel yaqkikyka  
sdeidhaid mtsi  
//
```

Upload a file:

COMS4761-- 2007

9

BLAST: Basic Local Alignment Search

- Altschul & Karlin [1990]; a family of algorithms
- Idea: find matches with significant score statistics
 - Find maximal segment pairs (MSP): segments with significant score
 - Based on extensive statistical theory (summarized soon)
- Base Algorithm:
 - Step 1: index DB for words of size W (W -mers);
index query sequence for W -mers with score $>$ Threshold
 - $W = 3$ for protein, 11 for nucleotides
 - Step 2: search for matches with high score (HSP=high scoring pairs)
 - Step 3: extend hits to maximal score segments
 - Step 4: report matches with score above S

COMS4761-- 2007

10

BLAST Step 1-3: Finding Short High-Scoring Pairs (HSP)

- Create an Index of W-mers for database & query
 - For proteins W=3 means a dictionary of $20^3=8000$ words
- Match W-mers that score above a threshold T
 - FASTA searches for exact matches of ktuples
 - BLAST, in contrast, searches for high scoring pairs (HSP)
 - Key idea: exploit the fast part of the search to max the score rather than push the maximization for later, slower, phases

Query: GSVEDTTGSQSLAALLNKCKT**POG**QRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

Neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSQ	13
PQA	12
PQN	12
etc...	

Neighborhood score threshold (T=13)

From A. Baxevanis: "Nucleotide and Protein Sequence Analysis I"
via Kellis & Indyk, MIT, "BLAST & Database Search, Lecture 2"
COMS4761-- 2007

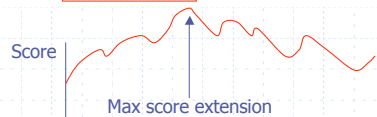
11

Blast Steps 3-4: Extending Short HSPs

- The short HSPs are extended to increase the score

Query: 325 SLAALLNKCKT**POG**QRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L+ TP G R++ +W+ P+ D + ER + A

Sbjct: 290 TLASVLDCTVTP**MG**SRMLKRWLHMPVRDTRVLLERQQTIGA 330



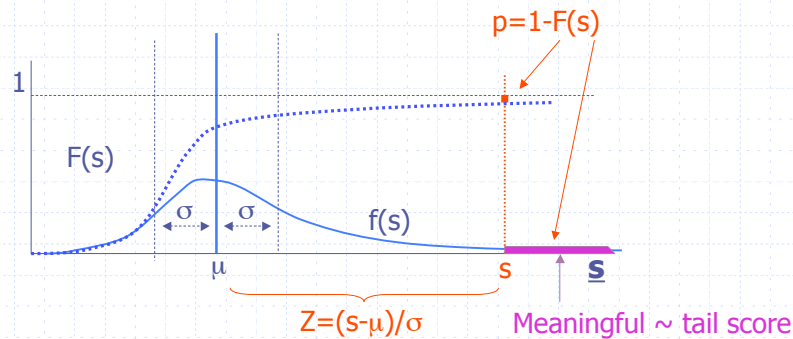
- Report above threshold HSPs and their scores

COMS4761-- 2007

12

Distinguishing Meaningful From Random

- “meaningful” ~ score is at the tail of the distribution
- Z-Score: $Z(s)=(s-\mu)/\sigma$ [$z(s) \geq 7 \rightarrow d$ is meaningful]
- P-Score: $p=p[\underline{S}>s]=1-F(s)$ [$p \leq 0.02 \rightarrow d$ is meaningful]



COMS4761-- 2007

15

How Significant is An Alignment?

- Key idea: Consider the highest matching score \underline{S} as a random variable and use Z-score to determine whether an alignment is meaningful
- Strategy:
 - Define the highest matching score \underline{S} as a random variable
 - Let q be a query and d a sequence from a database D
 - Define $\underline{S}(q,d)=\text{Max}\{s(q,q')\}$ where $s(q,q')$ is the score of an alignment of q and a subsequence q' of d
 - \underline{S} is a random variable defined over the space of local alignments $\{(q,d)\}$
 - Suppose \underline{S} has mean μ and standard deviation σ
 - Use Z-score, $Z(s)=(s-\mu)/\sigma$ to determine significance of a local alignment scoring s
 - For protein sequences $Z(s) > 7$ is considered significant
- Key challenge: how do we determine the distribution of \underline{S} ?

Karlin, S., and Altschul, S. F. (1990) "Method for assessing the statistical significance of molecular sequence features by using general scoring schemes," Proceedings of the National Academy of Science, USA 87, 2264-2268.

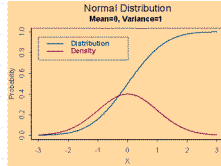
COMS4761-- 2007

16

What Is The Distribution of \underline{S} ?

- Consider the scores $\mathbf{S}'(q, q')$ for N random sequences q'
 - The score $\mathbf{S}'(q, q') = \sum_k s(q_k, q'_k)$ is the sum of independent random variables
 - For sufficiently long sequences $\underline{S}(q, q')$ is normally distributed

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$



- BLAST maximizes the score: $\underline{S}(q, d) = \text{Max}\{\mathbf{S}'(q, q') : q' \text{ is sub of } d\}$
 - \underline{S} has extreme order distribution: $F(x) = P[S < x] = P[\mathbf{S}'(q, q') < x \text{ for all } q' \text{ sub of } d]$
 - Extreme order theory: $F(x)$ is exponentially distributed
- Karlin-Altschul Statistics: $F(x) = P[\underline{S}(q, d) < x] \sim 1 - \exp[-(Kmn)\exp(-\lambda x)]$
 - Here $m=|q|$ $n=|d|$, λ and K may be computed from the scoring statistics
 - A good approximation for aligning sequences of length m, n is: $\lambda \sim \log(mn)$, $\sigma \sim 1$

COMS4761-- 2007

17

Application Example

- Consider local alignment of protein sequences (q, d)
 - Suppose the best local alignment is:


```
X'=FWLEVEGNSMTAPTG
Y'=FWLDVQGDSMTAPAG
```
- Compute the score of this local alignment:
 - Using PAM250, the score is $s=73$; normalized to log2 scale this gives $s'=\log_2(s)=24.3$
- Compute the parameters of a random scoring normal distribution:
 - Suppose $|q|=|d|=256$
 - The normal distribution for random scoring: $\mu=\log(mn)=\log(256*256)=16$, $\sigma=1$;
- Compute
 - $Z(s') = (24.3-16)/1 = 8.3 \text{ bits} > 7 \rightarrow$ alignment is significant

COMS4761-- 2007

18

Conclusions

- Indexing the sequence DB can accelerate alignment
 - FASTA: accelerate search for diagonal matches, then optimize alignment
 - BLAST: accelerate both, search for matches and optimizing the alignment

- FASTA innovations:
 - Exploit the geometry of diagonals
 - Grow alignments from seeds

- BLAST innovation: recast alignment as a pure search
 - Use scoring statistics as a mere guideline for the search
 - Contrast with DP use of (inexact) scoring to solve an exact optimization problem