# Genome 559: Introduction to Statistical and Computational Genomics

Winter 2009

Lecture14a

Multiple Sequence Alignment

Larry Ruzzo

# One Minute Responses

I like this pace; spot on (2)

Class was good today, easier to understand, examples were very useful, but helpful to have more examples of simple programs showing new concepts

I liked the blast part; blast material was great

Regexps are tricky; don't mind that we've spend so much time on them (2)

I liked the reg exp portion; good summary of past 2 classes

Struggling with reg exps, but think it will work itself out

I still appreciate the reiteration of concepts in class

Wish I'd seen list comprehension before previous hw.  Hope the new hw does not require a huge leap

How do "2-hit" & "gapped blast” heuristics relate to each other?

*Independent ideas, tho I think current BLAST does both by default*

I had some difficulty understanding the second threshold comparison when you explained BLAST.

*It's a multi-level filter. “Seed hits" below threshold1 get discarded. Above thresh1 they get extended (bidirectionally, ungapped) against the data base, and if the result is below thresh2, they also get discarded. Above that, they move on to the next steps, like gapped alignment.*

# Motivations

Common structure, function, or origin may be only weakly reflected in sequence; multiple comparisons may highlight weak signal

Major uses

- represent protein, RNA families
- represent & identify conserved seq features
- “whole genome” alignments

# Multiple Alignment Scoring

The Key Issue

Varying goals, methods (& controversy)

Ideal is perhaps phylogenetic, position specific, but typically too slow, too many parameters

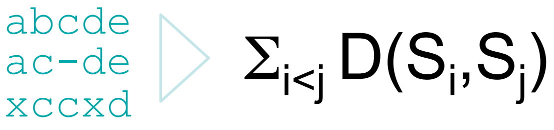
Most methods assume independence between columns, so you can score them separately

(Very inappropriate for RNA alignments, e.g.)

# Multiple Alignment Scoring within one column

Two common ways:

1. Min Entropy – if you assume a star phylogeny with long branches, positions in one column are independent and a proper probabilistic model reduces to per-column entropy (akin to last week). Intuitively sensible; favors alignments with less in-column variability
2. SP score: Sum of Pairs
   E.g., use BLOSUM62 score
   between all pairs of sequences

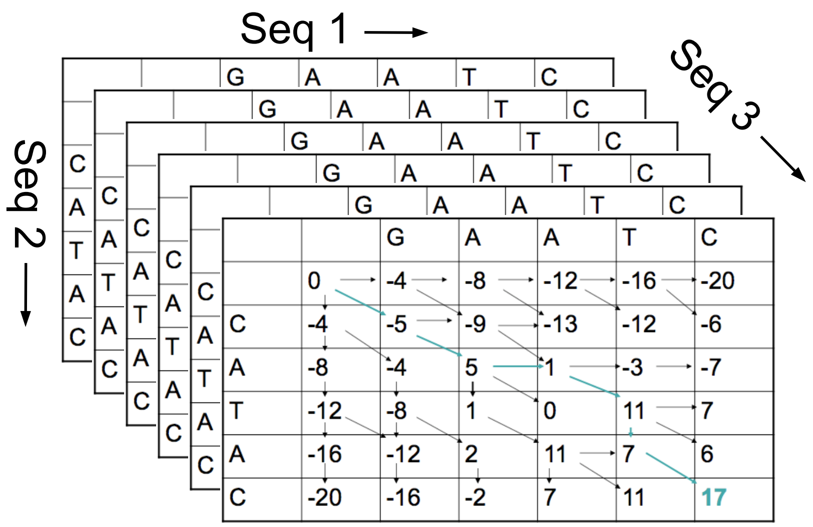


```
abcde
ac-de
xccxd
```

$\Sigma_{i<j} D(S_i,S_j)$

It is *not* theoretically justifiable, but is easy, not terrible

# Optimal SP Alignment via DP



| | | G | A | A | T | C |
|---|---|---|---|---|---|---|
| | 0 | -4 | -8 | -12 | -16 | -20 |
| C | -4 | -5 | -9 | -13 | -12 | -6 |
| A | -8 | -4 | 5 | 1 | -3 | -7 |
| T | -12 | -8 | 1 | 0 | 11 | 7 |
| A | -16 | -12 | 2 | 11 | 7 | 6 |
| C | -20 | -16 | -2 | 7 | 11 | 17 |

k sequences of length n

(n+1) x (n+1) x ··· x (n+1) k-dim array

Max of $2^k-1$ neighbors per cell; $(n+1)^k$ cells

Time: at least $(2n)^k$

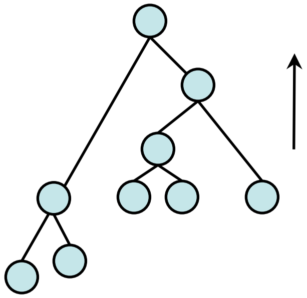
Want n, k 10's to 100's

Unlikely to do dramatically better – it's "NP-hard" Wang & Jiang, '94

E.g., n = 100
$10^6$ ops/sec

| k | Time |
|---|---|
| 2 | 40 ms |
| 3 | 8 sec |
| 4 | .5 hr |
| 5 | 100 hrs |
| 6 | 2 years |

# Common Heuristic: Progressive Alignment



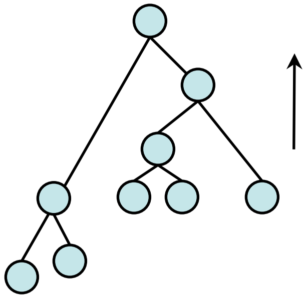
Pick a "guide tree"

- phylogeny would be ideal, but expensive
- quick alternative: get pairwise alignment scores, convert to distances, use, e.g., "neighbor joining"

Work up tree, leaves to root, doing pairwise alignments

(Many implementations, many variants, e.g. ClustalW)

# Aligning Alignments



Except at leaves, progressive alignment is aligning two alignments or a sequence to an alignment

Key in pairwise alignment is scoring “x aligns with y”

Now x, y are *columns* in the input alignments. Score?

Convenience of SP score is that you just score each letter in x vs each letter in y, say via BLOSUM62

Usual issues with gaps

Now run usual pairwise DP alignment

# Summary

Very important problem

Scoring is very difficult to get right

Fast, exact solutions appear impossible (even with simple scoring schemes)

Many heuristics have been tried

Useful methods like ClustalW are available

Still an open field

- e.g., “genome scale” and RNA especially challenging