# CSE 427
# Computational Biology

## Multiple Sequence Alignment

# Motivations

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

Major uses

- represent protein families
- represent & identify conserved seq features
- deduce evolutionary history

# Multiple Sequence Alignment

**Defn:** An *alignment* of $S_1, S_2, \ldots, S_k$,
is a set of strings $S'_1, S'_2, \ldots, S'_k$, (with spaces) s.t.

(1) $|S'_1| = |S'_2| = \ldots = |S'_k|$, and

(2) removing all spaces leaves $S_1, S_2, \ldots, S_k$

```
a c b c d b      a c – – b c d b
c a d b d        – c a d b – d –
a c a b c d      a c a – b c d –
```

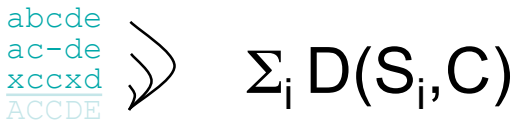
# Multiple Alignment Scoring

Varying goals

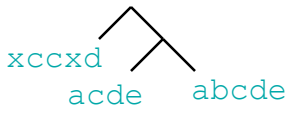
Varying methods (& controversy)

3 examples:

- Consensus string;
  sum distances to it




- Align to (evolutionary) tree;
  sum edges




- SP score:
  Sum of Pairs

# Optimal SP Alignment via DP

k strings 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-complete 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 |

# Center Star Alignment: A Bounded Error Approximation

Distance $\delta$, instead of similarity $\sigma$

Assume "Triangle Inequality":

$\delta(x,z) \leq \delta(x,y) + \delta(y,z)$

[plausible, but not always true]

Theorem: CSA gives MSA with SP score within 2 x of optimal

# Center Star Alignment: Method

D(S,T) = min distance of S-T alignment

Find $S_c$ minimizing $\sum_{i \neq c} D(S_c,S_i)$

For each unaligned string S

- Align $S'_c$ and S, giving $S''_c$ and S'
- Add new spaces in $S''_c$ to all previously aligned strings
- Add S' to set

# Center Star Alignment: Error Bound

I will completely skip proof, but it can be shown that this algorithm gives an answer that is within a factor of two of the optimal (under SP model).

2x comes from “Triangle Inequality”

# Center Star Alignment: Timing

Assume all strings of length n

$\binom{k}{2}$ pairwise alignments, $n^2$ each

$i^{th}$ addition costs (i*n)*n: $\Sigma_i\ in^2 = O(k^2n^2)$

Total time: $O(k^2n^2)$

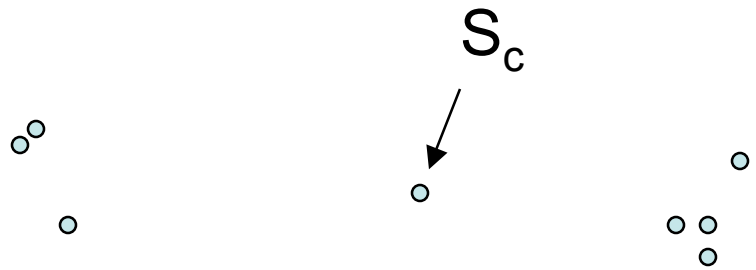
# Center Star Alignment: Notes

Error analysis doesn't mean it's always 2 x optimal

Better in practice and never worse

Could add "local optimizations" at end

# Where might CSA be poor?




Better to merge "clusters" first?

Why doesn't CSA do it?

-- Can't analyze it!

# Iterative Pairwise Alignment

Align some pair

While not done

- Pick an unaligned string “near” some aligned one(s)
- Align with the previously aligned group

Many variants

# Summarizing a Multiple Alignment

A *profile* of a multiple alignment gives letter frequencies per column

```
a b a
a b -
- b a
c a -
```

| | col 1 | col 2 | col 3 |
|---|---|---|---|
| a | 50% | 25% | 50% |
| b | 0% | 75% | 0% |
| c | 25% | 0% | 0% |
| - | 25% | 0% | 50% |

Alternatively, use log likelihood ratios

$p_i(a)$ = fraction of a's in col i

$p(a)$ = fraction of a's overall

$\log p_i(a)/p(a)$

# Aligning A String To A Profile

Key in pairwise alignment is scoring two positions x & y: $\sigma(x,y)$

For x a letter and y a (column) of a profile, let $\sigma(x,y)$ = value for x in col y

Invent a score for $\sigma(x,-)$

Run usual pairwise DP alignment

# Iterative Pairwise Alignment (More Detail)

align some pair

while not done

- Pick an unaligned string "near" some aligned one(s)
- Align with the profile of the previously aligned group
- Resulting new spaces inserted in all

Many variants

# Aligning to a Phylogenetic Tree

Given a tree with a sequence at each leaf, assign labels to internal nodes so as to minimize $\Sigma_{\text{edges (i,j)}} D(S_i,S_j)$

[Note: NOT SP score]

Also NP-Complete

Poly time approximation within 2 x possible; better with more time (PTAS)

# Progressive Alignment

Again, aligning to a tree

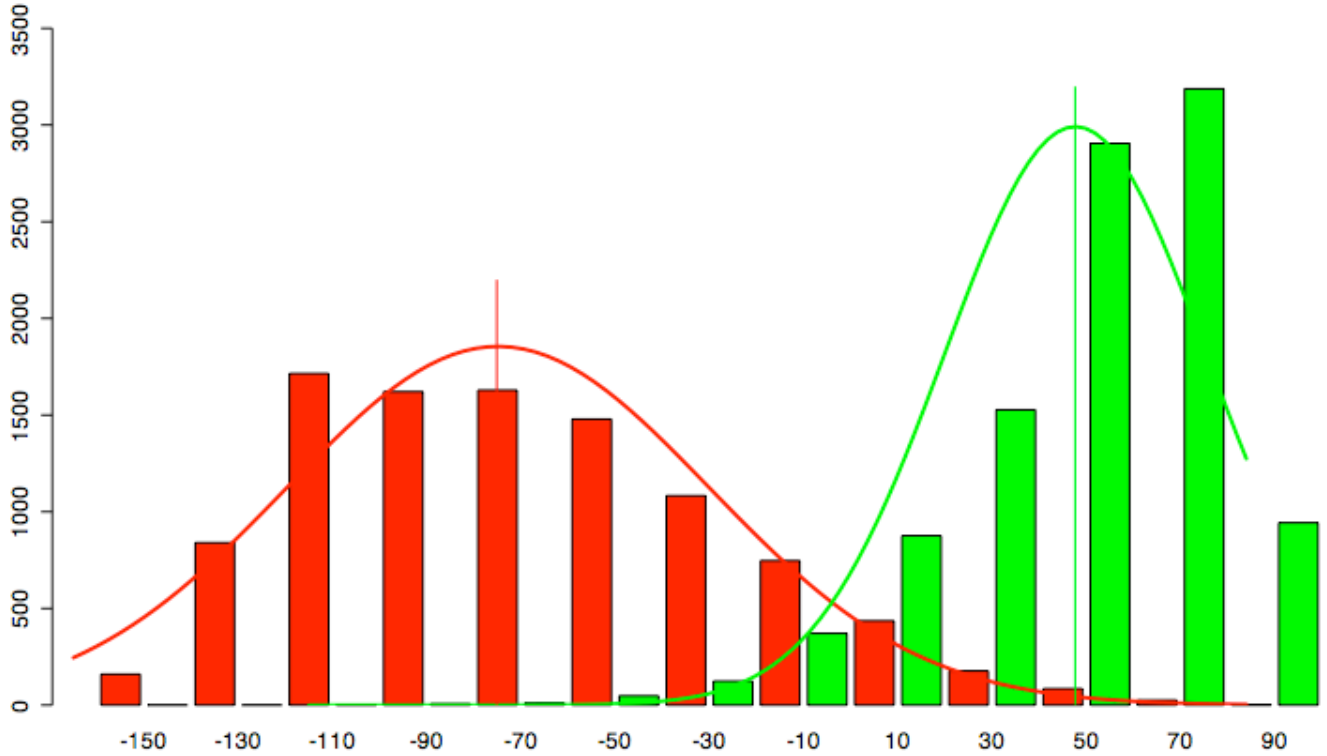
Initially, leaves labeled by strings; internal nodes unlabeled

at each step, pick an unlabeled node x with labeled children y,z

Align y & z, treating *columns* as units; give x that label

New feature: at general step, we're aligning two (smaller) alignments; score? (e.g. relative entropy)

# Similar Distributions?



# Relative Entropy

AKA Kullback-Liebler Distance/Divergence,
AKA Information Content

Given distributions P, Q

$$H(P||Q) = \sum_{x \in \Omega} P(x) \log \frac{P(x)}{Q(x)} \quad \geq 0$$

Notes:

Let $P(x) \log \frac{P(x)}{Q(x)} = 0$ if $P(x) = 0$ [since $\lim_{y \to 0} y \log y = 0$]

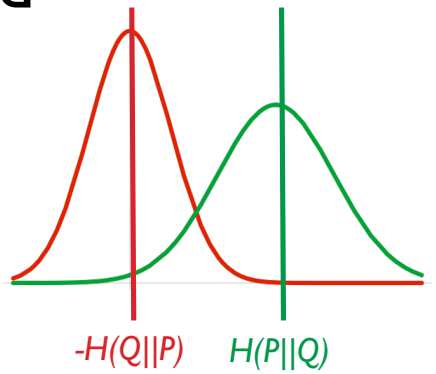
Undefined if $0 = Q(x) < P(x)$

# WMM: How "Informative"? Mean score of site vs bkg?

For any fixed length sequence *x*, let
*P(x)* = Prob. of *x* according to WMM
*Q(x)* = Prob. of *x* according to background

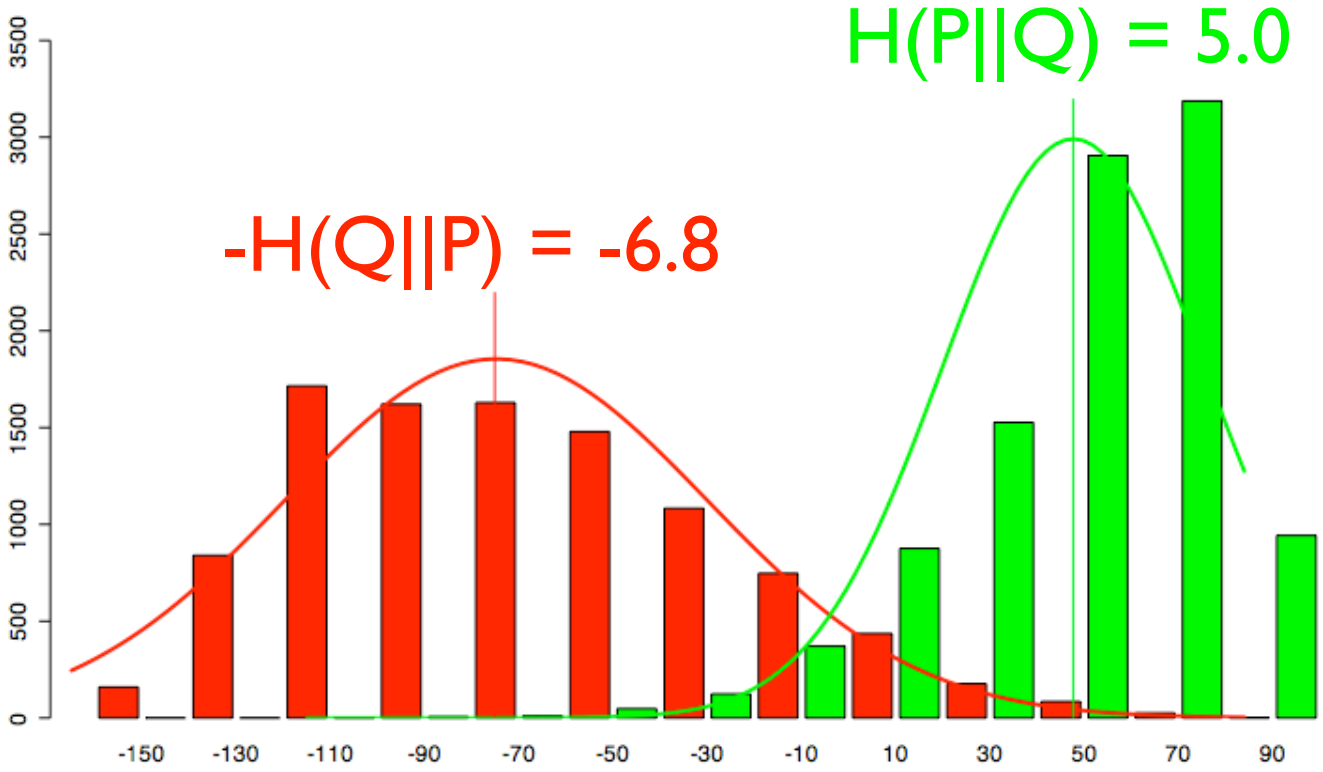
Relative Entropy:

$$H(P||Q) = \sum_{x \in \Omega} P(x) \log_2 \frac{P(x)}{Q(x)}$$




*H(P||Q)* is *expected log likelihood score* of a sequence randomly chosen from *WMM*;
*-H(Q||P)* is expected score of *Background*

# WMM Scores vs Relative Entropy



For WMM, you can show (based on the assumption of independence between columns), that :

$$H(P||Q) \quad = \quad \sum_i H(P_i||Q_i)$$

where $P_i$ and $Q_i$ are the WMM/background distributions for column i.

# Other Approaches

Other “spanning tree” algorithms

Other clustering algorithms

Repeated motifs

Hidden Markov Models

Gibbs sampling

...

# Summary

Very important problem

Exact solutions in poly time appear impossible

Bounded approximations are possible

Many heuristics have been tried

Still an open field