Parametric Sequence Alignment

Cynthia Vinzant

University of California, Berkeley

September 19, 2008

Introduction

Sequence Alignment

An Upper Bound

$\alpha - \beta$ plane

proof

$\sqrt{n}$ conjecture

A Lower Bound

Construction

Example for $q = 4$

Sequence Alignment

We have two sequences (representing species) and would like some measure their similarity.

Sequence Alignment

We have two sequences (representing species) and would like some measure their similarity.

To *align*, you insert spaces, to form new sequences.

Sequence Alignment

We have two sequences (representing species) and would like some measure their similarity.

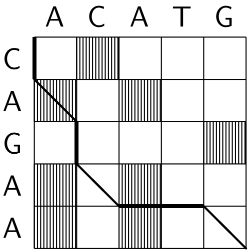
To *align*, you insert spaces, to form new sequences.

Example: One alignment of *ACTAG* and *CAGAA* is

$$\begin{array}{ccccccc} - & A & - & C & A & T & G \\ C & A & G & A & - & - & A \end{array}$$

Alignment Graphs

One way of representing an alignment of sequences of length n is as a path through a $n \times n$ grid.



$$\leftrightarrows \qquad \begin{array}{ccccccc} - & A & - & C & A & T & G \\ C & A & G & A & - & - & A \end{array}$$

Alignment Summaries

Every alignment has an *alignment summary*, (w, x, y), where

- $w = \#$ of matches
- $x = \#$ of mismatches
- $y = \#$ of spaces

Alignment Summaries

Every alignment has an *alignment summary*, (w, x, y), where

- $w = \#$ of matches
- $x = \#$ of mismatches
- $y = \#$ of spaces

Example:

$$\begin{array}{ccccccc} - & A & - & C & A & T & G \\ C & A & G & A & - & - & A \end{array} \quad \longrightarrow \quad (1,2,2)$$

Alignment Summaries

Every alignment has an *alignment summary*, (w, x, y), where

- $w = \#$ of matches
- $x = \#$ of mismatches
- $y = \#$ of spaces

Example:

$$\begin{array}{cccccccc} - & A & - & C & A & T & G \\ C & A & G & A & - & - & A \end{array} \quad \longrightarrow \quad (1,2,2)$$

**NOTE: $w + x + y = n$ ($=$ length of sequences)

Optimal Alignments of ACATG, CAGAA

Alignments		*Alignment Summaries*
Γ_1 :	CAGAA ACATG	$(0, 5, 0)$
Γ_2 :	−CAGAA ACATG−	$(2, 2, 1)$
Γ_3 :	−CA−GAA ACATG−−	$(3, 0, 2)$
Γ_4 :	−−−−−CAGAA ACATG−−−−−	$(0, 0, 5)$

How many optimal alignment summaries are there?

Theorem (Gusfield, 1994)

For any alphabet Σ, $f_\Sigma(n) = O(n^{2/3})$, that is, there is a constant c s.t.

$$f_\Sigma(n) \leq c \cdot n^{2/3}$$

How many optimal alignment summaries are there?

Theorem (Gusfield, 1994)

For any alphabet Σ, $f_\Sigma(n) = O(n^{2/3})$, that is, there is a constant c s.t.

$$f_\Sigma(n) \leq c \cdot n^{2/3}$$

Theorem (Fernández-Baca et. al., 2002)

$$f_\Sigma(n) \leq \frac{3}{(2\pi)^{2/3}} n^{2/3} + O(n^{1/3} \log n)$$

Observations

1) Boundary lines are of the form

$$\{(\alpha, \beta) : \mathsf{score}_{\alpha,\beta}(\Gamma_1) = \mathsf{score}_{\alpha,\beta}(\Gamma_2)\}.$$

Observations

1) Boundary lines are of the form

$$\{(\alpha,\beta) : \text{score}_{\alpha,\beta}(\Gamma_1) = \text{score}_{\alpha,\beta}(\Gamma_2)\}.$$

2) All boundary lines pass through the point $(-1,-1)$.

Observations

1) Boundary lines are of the form

$$\{(\alpha, \beta) : \mathrm{score}_{\alpha,\beta}(\Gamma_1) = \mathrm{score}_{\alpha,\beta}(\Gamma_2)\}.$$

2) All boundary lines pass through the point $(-1, -1)$.

3) No boundary lines intersect the ray $\beta = 0$, $\alpha > 0$.

Observations

1) Boundary lines are of the form

$$\{(\alpha, \beta) : \text{score}_{\alpha,\beta}(\Gamma_1) = \text{score}_{\alpha,\beta}(\Gamma_2)\}.$$

2) All boundary lines pass through the point $(-1, -1)$.

3) No boundary lines intersect the ray $\beta = 0$, $\alpha > 0$.

Conclusion: All boundary lines must (uniquely) intersect the non-negative β-axis.

Simplification

Now we only need to know how many optimality regions there are on the non-negative β-axis.

Simplification

Now we only need to know how many optimality regions there are on the non-negative β-axis.

Boundary lines (now just points) look like

$$\{\beta : \text{score}_{(0,\beta)}(\Gamma_1) = \text{score}_{(0,\beta)}(\Gamma_2)\},$$

Simplification

Now we only need to know how many optimality regions there are on the non-negative β-axis.

Boundary lines (now just points) look like

$$\{\beta : \text{score}_{(0,\beta)}(\Gamma_1) = \text{score}_{(0,\beta)}(\Gamma_2)\},$$

meaning

$$w_1 - \beta y_1 = w_2 - \beta y_2 \quad \Rightarrow \quad \beta = \frac{w_2 - w_1}{y_2 - y_1}$$

Constraints

We only need to look at (w, y)-plane.

- Suppose we have vertices $(w_1, y_1), (w_2, y_2), \ldots, (w_m, y_m)$ of an alignment polytope for sequences of length n.
  (Want to know: How big can m be in terms of n?)

Constraints

We only need to look at (w, y)-plane.

- Suppose we have vertices $(w_1, y_1), (w_2, y_2), \ldots, (w_m, y_m)$ of an alignment polytope for sequences of length n.
  (Want to know: How big can m be in terms of n?)
- Note:

$$\sum_{i=1}^{m} \Delta w_i \leq n \qquad \text{and} \qquad \sum_{i=1}^{m} \Delta y_i \leq n$$

Constraints

We only need to look at (w, y)-plane.

- Suppose we have vertices $(w_1, y_1), (w_2, y_2), \ldots, (w_m, y_m)$ of an alignment polytope for sequences of length n.
  (Want to know: How big can m be in terms of n?)

- Note:

$$\sum_{i=1}^{m} \Delta w_i \leq n \qquad \text{and} \qquad \sum_{i=1}^{m} \Delta y_i \leq n$$

- To maximize m, we need the most distinct $\frac{\Delta w_i}{\Delta y_i}$, with $\Delta w_i + \Delta y_i$ small.

Given that $\sum_i (\Delta w_i + \Delta y_i) \leq 2n$, how many different slopes $\frac{w_i}{y_i}$ can there be?

Given that $\sum_i (\Delta w_i + \Delta y_i) \leq 2n$, how many different slopes $\frac{w_i}{y_i}$ can there be?

Define $F_r = \{\frac{a}{b} \quad \text{s.t.} \quad a + b = r \text{ and } a, b \text{ relatively prime}\}$

Given that $\sum_i(\Delta w_i + \Delta y_i) \leq 2n$, how many different slopes $\frac{w_i}{y_i}$ can there be?

Define $F_r = \{\frac{a}{b} \quad \text{s.t.} \quad a + b = r$ and a, b relatively prime$\}$

Taking our $\frac{\Delta w_i}{\Delta y_i}$ from $\bigcup_{r=1}^{q} F_r$ gives us

$$m = \sum_{r=1}^{q} |F_r| \quad \text{and} \quad n = \sum_{r=1}^{q} r \cdot |F_r|$$

Given that $\sum_i(\Delta w_i + \Delta y_i) \leq 2n$, how many different slopes $\frac{w_i}{y_i}$ can there be?

Define $F_r = \{\frac{a}{b} \quad \text{s.t.} \quad a + b = r \text{ and } a, b \text{ relatively prime}\}$

Taking our $\frac{\Delta w_i}{\Delta y_i}$ from $\bigcup_{r=1}^{q} F_r$ gives us

$$m = \sum_{r=1}^{q} |F_r| \qquad \text{and} \qquad n = \sum_{r=1}^{q} r \cdot |F_r|$$

$\Rightarrow$

$$m \approx q^2 \qquad \text{and} \qquad n \approx q^3$$

Can the bound be improved for finite alphabets?

Fernández-Baca et. al. (2002) showed

- $c \cdot n^{2/3} \leq f_{\Sigma}(n)$ for Σ infinite (by constructing sequences that attained $n^{2/3}$ optimal alignments).
- $c \cdot \sqrt{n} \leq f_{\{0,1\}}(n)$.
- $E(g(\sigma_1, \sigma_2) \approx \sqrt{n}$.

Can the bound be improved for finite alphabets?

Fernández-Baca et. al. (2002) showed

- $c \cdot n^{2/3} \leq f_\Sigma(n)$ for Σ infinite (by constructing sequences that attained $n^{2/3}$ optimal alignments).
- $c \cdot \sqrt{n} \leq f_{\{0,1\}}(n)$.
- $E(g(\sigma_1, \sigma_2)) \approx \sqrt{n}$.

Conjecture: $f_{\{0,1\}}(n) = \Theta(\sqrt{n})$. That is, there are constants, c, C so that

$$c \cdot \sqrt{n} \leq f_{\{0,1\}}(n) = C \cdot \sqrt{n}$$

The bound is tight!

Claim: $f_{\{0,1\}}(n) = \Theta(n^{2/3})$.

So our goal is to construct two sequences σ_1, σ_2 that have $n^{2/3}$ optimal alignments.

Define

$$\overline{F}_r = \{\frac{a}{b} \leq 1 \text{ s.t. } a, b \text{ relatively prime, and } a + b = r\},$$

and let $\{\frac{a_1}{b_1}, \frac{a_2}{b_2}, \dots, \frac{a_m}{b_m}\}$ be the elements of $\bigcup_{r=1}^{q} \overline{F}_r$.

Define

$$\overline{F}_r = \{\frac{a}{b} \leq 1 \text{ s.t. } a, b \text{ relatively prime, and } a + b = r\},$$

and let $\{\frac{a_1}{b_1}, \frac{a_2}{b_2}, \ldots, \frac{a_m}{b_m}\}$ be the elements of $\displaystyle\bigcup_{r=1}^{q} \overline{F}_r$.

First sequence, $\sigma_1 =$

$$0^{b_1+a_1} 1^{b_1-a_1} 0^{b_2+a_2} \ldots 0^{b_m+a_m} 1^{b_m-a_m} \; 0^{b_m-a_m} 1^{b_m+a_m} \ldots 0^{b_1-a_1} 1^{b_1+a_1}$$

Define

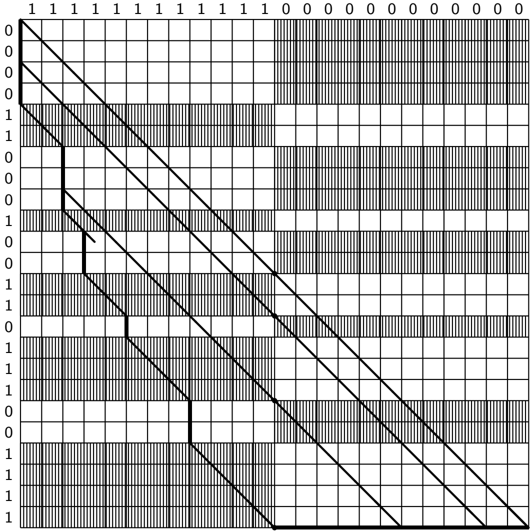
$$\overline{F}_r = \{\frac{a}{b} \leq 1 \text{ s.t. } a, b \text{ relatively prime, and } a + b = r\},$$

and let $\{\frac{a_1}{b_1}, \frac{a_2}{b_2}, \ldots, \frac{a_m}{b_m}\}$ be the elements of $\bigcup_{r=1}^{q} \overline{F}_r$.

First sequence, $\sigma_1 =$

$$0^{b_1+a_1}1^{b_1-a_1}0^{b_2+a_2}\ldots 0^{b_m+a_m}1^{b_m-a_m}\ \ 0^{b_m-a_m}1^{b_m+a_m}\ldots 0^{b_1-a_1}1^{b_1+a_1}$$

Second sequence, $\sigma_2 = 1^{(\Sigma 2b_i)}0^{(\Sigma 2b_i)}$.


1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
0
0
0
0
1
1
0
0
0
1
0
0
1
1
0
1
1
1
0
0
1
1
1
1

Some open questions:

1) What is $E(g(\sigma_1, \sigma_2))$? $\Theta(\sqrt{n})$?

2) Pachter and Sturmfels (year) showed that for d-parameter models, $f_\Sigma(n) = O(n^{d(d-1)/(d+1)})$. Is this also tight?