The Role of Convex Optimization in Optimal Alignments of Random Sequences

Convex Optimization and Beyond

Friday 27 June 2014, ICMS Edinburgh

Raphael Hauser, Oxford Mathematical Institute and Pembroke College

Based on collaborations with Saba Amsalu, Clément Durringer, Servet Martinez and Heinrich Matzinger. Supported by EPSRC grants EP/H02686X/1 and EP/I01893X/1, IMA Grant SGS29/11

Why Study Sequence Alignments?

- Where does a finite sequence fit best in a larger finite sequence?

$$gattggatcctagccctgagagatttccccccctaggaaaatttcat\ldots$$
$$\leftarrow actg \rightarrow$$

- What finite sequence among a finite choice of candidates is a best fit for a given new sequence?

$$0100011 \stackrel{?}{\rightarrow} \begin{cases} 0100001110101 \\ 11010110 \\ 111111011111 \end{cases}$$

How to Align Sequences?

$$\begin{array}{cccccccc} b & g & \star & h & o & u & s & e \\ b & i & g & \star & h & o & s & e \end{array}$$

$$\begin{array}{cccccccc} b & \sqcup & g & \star & h & o & u & s & e \\ b & i & g & \star & h & o & \sqcup & s & e \end{array}$$

- Allow the introduction of gaps subject to a penalty.
- Assign a similarity score to each pair of aligned letters.
- Maximise total score over the set of possible alignments.

To align $x_1x_2x_3 \in \mathcal{A}^3$ with $y_1y_2y_3y_4 \in \mathcal{A}^4$:

- Fix a scoring function $s : \mathcal{A}^* \times \mathcal{A}^* \to \mathbb{R}$.

- E.g., for the alignment

$$\pi : \begin{matrix} \mathfrak{G} & x_1 & x_2 & x_3 & \mathfrak{G} \\ y_1 & y_2 & \mathfrak{G} & y_3 & y_4 \end{matrix}$$

  with gaps $\mathfrak{G}$, define the total alignment score by

$$S_\pi(x, y) = s(\mathfrak{G}, y_1) + s(x_1, y_2) + s(x_2, \mathfrak{G}) + s(x_3, y_3) + s(\mathfrak{G}, y_4).$$

- To find an optimal alignment, solve

$$L_s(x, y) = \max_\pi S_\pi(x, y),$$
$$\pi^* = \arg\max_\pi S_\pi(x, y)$$

Example 1. LCS scoring function

$$
\begin{aligned}
s(\mathfrak{a}, \mathfrak{b}) &= -1 \quad \text{if } \mathfrak{a} \neq \mathfrak{b}, \\
s(\mathfrak{a}, \mathfrak{a}) &= 1 \quad \forall\, \mathfrak{a} \in \mathcal{A}, \\
s(\mathfrak{G}, \mathfrak{a}) &= s(\mathfrak{a}, \mathfrak{G}) = 0 \quad \forall\, \mathfrak{a} \in \mathcal{A}.
\end{aligned}
$$

- $L_s(x, y)$ is the length of a longest common subsequence of x and y.

- π^* identifies a longest common subsequence by dropping letters aligned with a gap in both sequences.

Example 2. BLASTZ scoring function for $\mathcal{A} = \{A, T, C, G\}$:

	A	T	C	G	$\mathfrak{G}$
A	91	-31	-114	-123	-400
T	-31	100	-125	-114	-400
C	-114	-125	100	-31	-400
G	-123	-114	-31	91	-400
$\mathfrak{G}$	-400	-400	-400	-400	$-\infty$

- $s(\mathfrak{a}, \mathfrak{b})$ represents log-likelihood that $\mathfrak{a}$ evolved into $\mathfrak{b}$ under a stochastic evolutionary model.

- $L_s(x, y)$ yields the maximum log-likelihood that x and y arose from a common ancestor via mutation, and the most likely microstructure of the mutations is exhibited by π^*.

- In reality, the log-likelihood depends on the time since evolutionary divergence.

Algorithm 1 (Dynamic Programming)**.**

1. $L_s(\emptyset, \emptyset) = 0$

   for $i, j \geq 1$

   $L_s(x_{[1,i]}, \emptyset) = \sum_{\ell=1}^{i} s(x_\ell, \mathfrak{G})$

   $L_s(\emptyset, y_{[1,j]}) = \sum_{\ell=1}^{j} s(\mathfrak{G}, y_\ell)$

   end

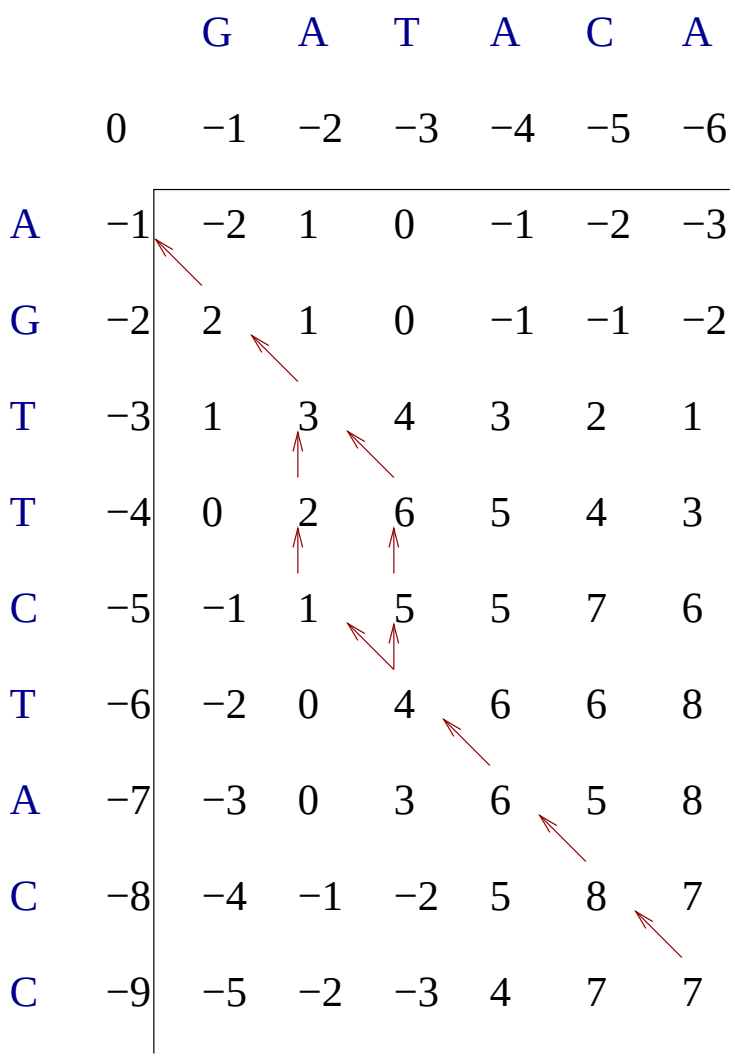
2. for $i = 1, \ldots \text{length}(x)$

   for $j = 1, \ldots \text{length}(y)$

   $$L_s(x_{[1,i]}, y_{[1,j]}) = \max\big(s(x_i, y_j) + L_s(x_{[1,i-1]}, y_{[1,j-1]}),$$
   $$s(x_i, \mathfrak{G}) + L_s(x_{[1,i-1]}, y_{[1,j]}), s(\mathfrak{G}, y_j) + L_s(x_{[1,i]}, y_{[1,j-1]})\big)$$

   end

   end


G A T A C A
0 −1 −2 −3 −4 −5 −6
A −1 −2 1 0 −1 −2 −3
G −2 2 1 0 −1 −1 −2
T −3 1 3 4 3 2 1
T −4 0 2 6 5 4 3
C −5 −1 1 5 5 7 6
T −6 −2 0 4 6 6 8
A −7 −3 0 3 6 5 8
C −8 −4 −1 −2 5 8 7
C −9 −5 −2 −3 4 7 7

Understanding the Null-Model

To decide on the significance of high scores, compare with scores achieved by random sequences

$$L_{m,n}(s) = L_s(X_{[1,m]}, Y_{[1,n]}) = L_s(X_1 X_2 \dots X_m\,,\, Y_1 \dots Y_n).$$

with i.i.d. letters $X_i, Y_j : \Omega \to \mathcal{A}$.

When $m = n$ we write $L_n(s) = L_{m,n}(s)$.

Our understanding of this null-model is still incomplete.

Three broad themes with analogue questions in percolation theory:

- Determine the Chvàtal-Sankoff constant

$$\lambda(s,\xi) = \lim_{n\to\infty} \lambda_n(s,\xi)$$

  where $\xi \in [0,1]$ is fixed and

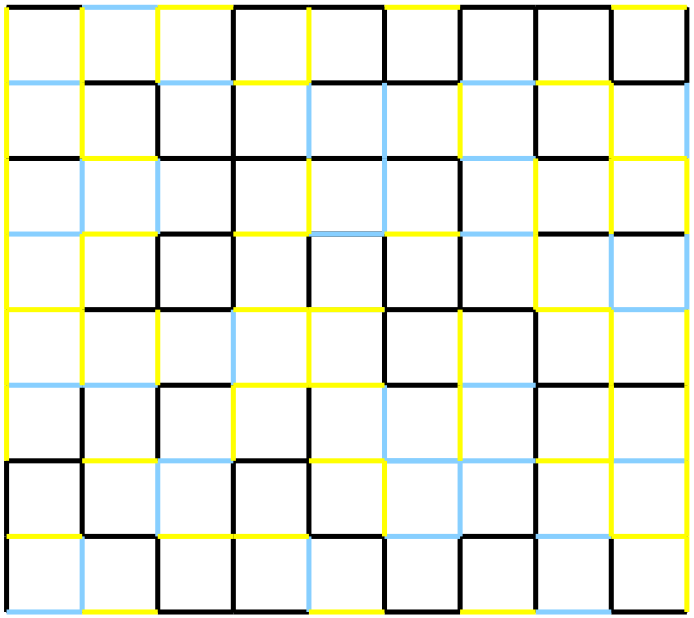
$$\lambda_n(s,\xi) = \frac{1}{n}\,\mathsf{E}\left[L_{\lfloor 2n\xi\rfloor,\lfloor 2n(1-\xi)\rfloor}(s)\right].$$

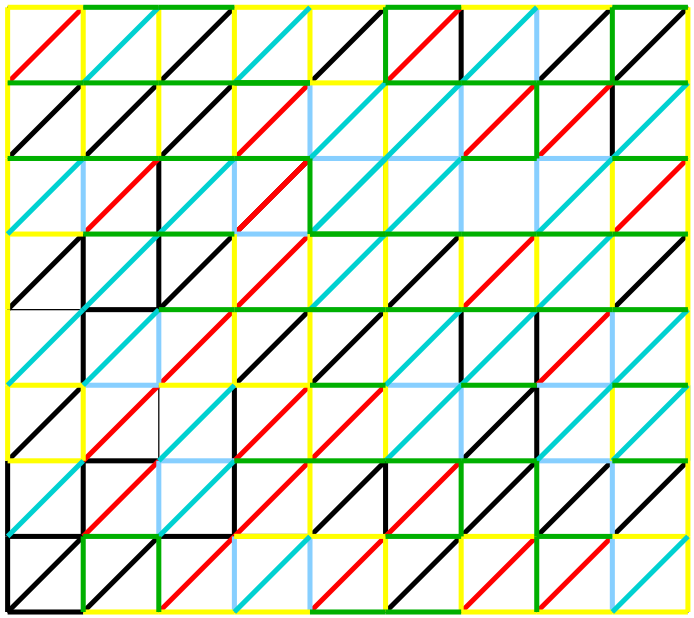
  We write $\lambda(s) = \lambda(s,1/2)$, $\lambda_n(s) = \lambda_n(s,1/2)$, and $(\xi_1,\xi_2) = (\xi, 1-\xi)$.

- Determine the fluctuation order

$$\mathsf{STD}\left(L_{\lfloor 2n\xi_1\rfloor,\lfloor 2n\xi_2\rfloor}(s)\right).$$

- Understand the microstructure of optimal alignments.





The Chvàtal-Sankoff Constant

- Subadditivity yields $\lambda_n(s) \nearrow \lambda(s)$ and in fact, $L_n(s)/n \to \lambda(s)$ almost surely.

- However, the convergence is very slow

$$\sqrt{2}\|s\|_\delta \frac{\sqrt{\ln n}}{\sqrt{n}} + \frac{2\|s\|_\infty}{n} \leq \lambda(s) - \lambda_n(s) \leq C\frac{\sqrt{\ln n}}{\sqrt{n}},$$

  where

$$\|s\|_\delta = \max_{\mathfrak{c},\mathfrak{d},\mathfrak{e}\in\mathcal{A}^*} |s(\mathfrak{c},\mathfrak{d}) - s(\mathfrak{c},\mathfrak{e})|\,,$$
$$\|s\|_\infty = \max_{\mathfrak{c},\mathfrak{d}\in\mathcal{A}^*} |s(\mathfrak{c},\mathfrak{d})|\,.$$

- Upper bound due to K. Alexander ["The rate of convergence of the mean length of the longest common subsequence." Ann. Appl. Prob. 4(4):1074–1082, 1994].

Example 3. For the LCS-scoring function, we have

$$\|s\|_\infty = 1,$$
$$\|s\|_\delta = 2.$$

- To obtain $|\lambda(s) - \lambda_n(s)| \leq 1\,\mathrm{e}\,{-2}$, we would need

$$n \approx 1.8\,\mathrm{e}\,6.$$

- To obtain $|\lambda(s) - \lambda_n(s)| \leq 1e - 3$, we would need

$$n \approx 2.4\,\mathrm{e}\,8$$

- The computational cost is $\mathrm{O}(n^2)$.

Accelerated Montecarlo Simulation for LCS

Parse sequences into pieces with known LCS,

$$\star\star\star\star\star\circ\circ\circ\star\star\star\star\circ\circ\circ\circ\circ\circ\circ\circ\circ\circ\circ\ldots$$
$$\star\star\star\star\circ\circ\circ\circ\circ\star\star\star\star\star\star\star\circ\circ\ldots$$

Two sequences $x^1 = x^1_1 \ldots x^1_{n_1}$, $x^2 = x^2_1 \ldots x^2_{n_2}$ are an m-match if

$$\begin{aligned} &L_s(x^1, x^2) = m, \\ &L_s(x^1_{[1,n_1-1]}, x^2) = m-1, \\ &L_s(x^1, x^2_{[1,n_2-1]}) = m-1, \end{aligned}$$

that is, removing the last character of any of the sequences reduces the LCS-score.

Lemma 2. *$L_n(\xi) \geq qn$ can only occur if*

$$\begin{bmatrix} x_1^1 \dots x_{\lfloor 2n\xi_1 \rfloor}^1 \\ x_1^r \dots x_{\lfloor 2n\xi_2 \rfloor}^2 \end{bmatrix} = \begin{bmatrix} x_1^1 \dots x_{\varsigma^1(m)}^1 \\ x_1^2 \dots x_{\varsigma^2(m)}^2 \end{bmatrix} \circ \dots \circ \begin{bmatrix} x_{\varsigma^1((k-1)m)+1}^1 \cdots x_{\varsigma^1(km)}^1 \\ x_{\varsigma^2((k-1)m)+1}^2 \cdots x_{\varsigma^2(km)}^2 \end{bmatrix} \circ \begin{bmatrix} \dots \\ \vdots \\ \dots \end{bmatrix}$$

is a concatenation of $k := \lfloor qn/m \rfloor$ m-matches and a remainder.

Corollary 3.

$$\mathsf{P}[L_n(\xi) \geq nq] \leq \left(\nu^{m,2}\right)^{*k} ([nq, \lfloor 2n\xi_1 \rfloor] \times [nq, \lfloor 2n\xi_2 \rfloor]) \,,$$

where $\left(\nu^{m,2}\right)^{*k}$ *is the* k*-fold convolution of the measure* $\nu^{m,2}$ *defined on* $\mathbb{N}^2$ *by*

$$\nu^{m,2}(i_1, i_2) := \mathsf{P}\left[(X_1^1 \ldots X_{i_1}^1 \,,\, X_1^2 \ldots X_{i_2}^2) \;\; \textit{is a } m\textit{-match}\right].$$

Let $B := \left[m \,,\, 2m\xi_1/q\right] \times \left[m \,,\, 2m\xi_2/q\right]$, so that

$$kB \approx [nq, \lfloor 2n\xi_1 \rfloor] \times [nq, \lfloor 2n\xi_2 \rfloor].$$

Theorem 4. *The following are equivalent,*

i) $\lambda(s,\xi) \le q$,

ii) $\limsup_{k\to\infty}\Big((\nu^{m,2})^{*k}(kB)\Big)^{1/k} < 1$,

iii) $\inf\Big\{\Lambda^{m,2}(x) - \frac{2m}{q}\langle\xi,x\rangle : \ x\in\mathbb{R}^2_-\Big\} \ < \ 0$, *where*

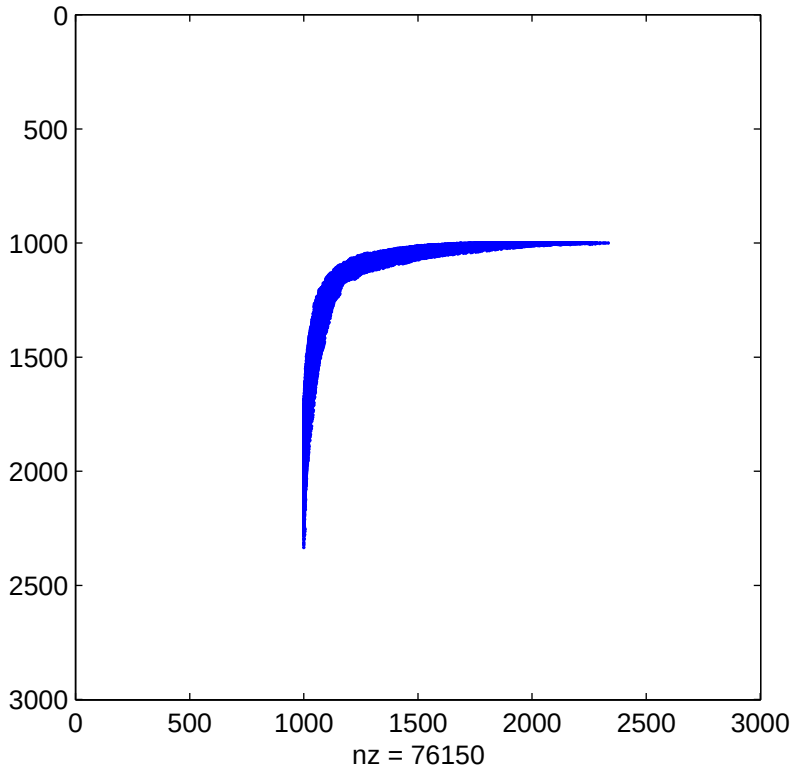
$$\Lambda^{m,2}(x) := \log\int_{\mathbb{R}^2}\nu^{m,2}(y)\,\mathrm{e}^{\langle y,x\rangle}\,dy$$

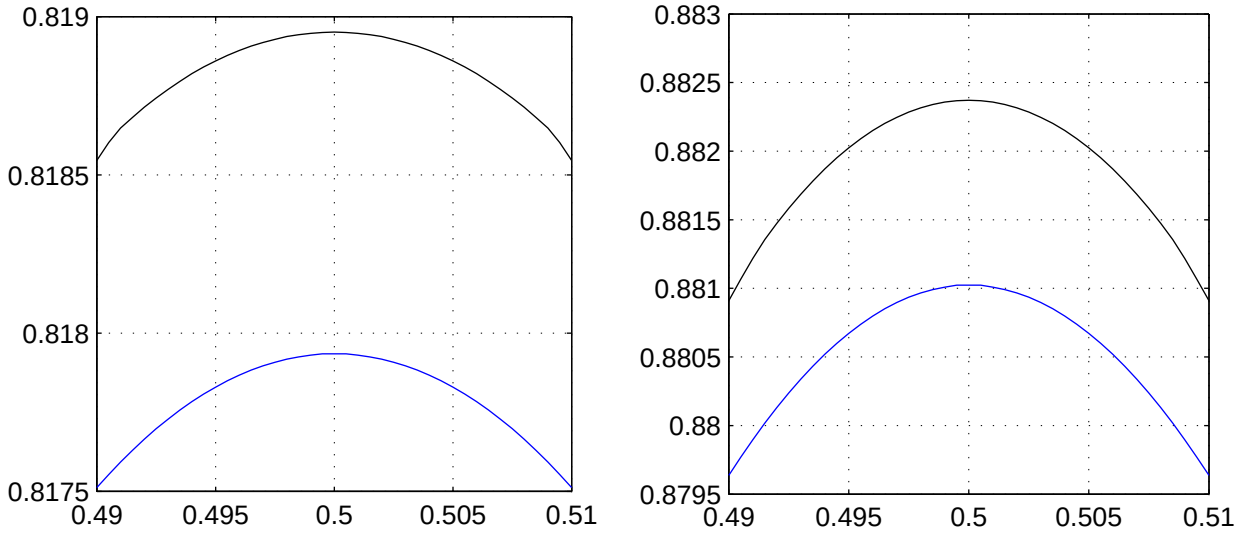
is the log-Laplace transform of $\nu^{m,2}$.

Example 4 (Durringer-H.-Matzinger, 2008)**.**

- Experiment 1: $\mathcal{A} = \{0, 1\}$, $P[X_i = 0] = 1/2$, $m = 1,000$.
- Experiment 2: $\mathcal{A} = \{0, 1\}$, $P[X_i = 0] = 0.2$, $m = 1,000$.
- Confidence level 95%.
- m and the confidence level were chosen such that computational cost equals a single simulation of $L_n(s)$ with $n = 1\,\mathrm{e}\,5$.

Experiment 2: Sparsity pattern of $\widehat{\nu}$ after $10,000$ simulations.





Estimated confidence intervals for $\lambda(s, \xi)$ show that the achieved accuracy is approximately $2\,\mathrm{e}\,{-3}$.

This corresponds to a 300'000-fold (!!) reduction in flops in comparison to brute-force simulation, which would require $n \approx 5.4\,\mathrm{e}\,7$ to achieve the same accuracy.

Steele conjecture [M.J. Steele. "An Effron-Stein inequality for non-symmetric statistics." Ann. Stat., 14: 753-758, 1958]: For LCS with i.i.d. U($\mathcal{A}$)-distributed variables X_i, Y_j,

$$\lambda(s, 0.5) = \frac{2}{1 + \sqrt{|\mathcal{A}|}}.$$

But e.g. for $|\mathcal{A}| = 2$,

$$\frac{2}{1 + \sqrt{|\mathcal{A}|}} = 0.8284 > 0.8182,$$

so with high confidence the Steele conjecture is wrong [H.-Martinez-Matzinger, 2006].

The Order of Fluctuations

Let $X = X_1 \dots X_n$, $Y = Y_1 \dots Y_n$ with i.i.d. Ber(p) letters.

- Chvàtal-Sankoff ["Longest common subsequence of two random sequences", J. Appl. Prob., 12:306–315, 1975] conjectured that in the case $p = 1/2$,

$$\mathrm{VAR}(L_n(s)) = \mathrm{o}\left(n^{2/3}\right).$$

- Steele ["An Effron-Stein inequality for non-symmetric statistics", Ann. Stat. 14:753–758, 1986] proved

$$\mathrm{VAR}(L_n(s)) \leq 2p(1-p)n.$$

- Waterman ["Estimating statistical significance in sequence alignment", Phil. Trans. R. Soc. Lond. B, 344:383-390, 1994] reports simulations that suggest that for $p < 0.5$,

$$\mathrm{VAR}(L_n(s)) = \Theta(n).$$

- Boutet de Monvel ["Extensive simulations for longest common subsequences." Eur. Phys. J. B, 7:293–308, 1999] reports simulations that suggest taht for $p = 0.5$,

$$\mathrm{VAR}(L_n(s)) = \Theta(n).$$

More recently, H.-Matzinger [2005], Lember-Matzinger [2009], Amsalu-H.-Matzinger [2012] showed that if

$$\mathsf{E}\left[L_s(\tilde{X}_{[1,n]}, \tilde{Y}_{[1,n]}) - L_s(X_{[1,n]}, Y_{[1,n]}) \,\|\, X, Y\right] \geq c$$

holds with probability $1 - \mathsf{O}(n^{-\alpha n})$, then $\mathsf{VAR}(L_n(s)) = \Theta(n)$.

- True for arbitrary scoring function s, alphabet $\mathcal{A}$, and distribution of X_i, Y_i.

- $\tilde{X}, \tilde{Y}$ obtained by selecting one letter of a specified type $\mathfrak{a}$ uniformly at random from the realised letters $X(\omega), Y(\omega)$ and changing it into a specified other type $\mathfrak{b}$.

- Bias analytically provable only in highly assymetric cases.

- Deeply connected with the problem of understanding the microstructure of optimal alignments.

- Basic tool used is Azuma-Hoeffding Inequality in the form given by McDiarmid ["On the method of bounded differences." Surveys in Combinatorics, 141:148–188, 1989]:

Theorem 5. *Let $Z_1, Z_1, \ldots, Z_m$ be i.i.d. random variables that take values in a set D, and let $g : D^m \to \mathbb{R}$ be a function of m variables with the property that*

$$\max_{i=1,\ldots,m} \sup_{z \in D^m, \widehat{z}_i \in D} |g(z_1, \ldots, z_m) - g(z_1, \ldots, \widehat{z}_i, \ldots, z_m)| \leq C.$$

Thus, changing a single argument of g changes its image by less than a constant C. Then the following bounds hold,

$$\mathsf{P}\left[g(Z_1, \ldots, Z_m) - \mathsf{E}[g(Z_1, \ldots, Z_m)] \geq \epsilon \times m\right] \leq \exp\left\{-\frac{2\epsilon^2 m}{C^2}\right\},$$

$$\mathsf{P}\left[\mathsf{E}\left[g(Z_1, \ldots, Z_m)\right] - g(Z_1, \ldots, Z_m) \geq \epsilon \times m\right] \leq \exp\left\{-\frac{2\epsilon^2 m}{C^2}\right\}.$$

Alignment Microstructure

Let π be an alignment with gaps of two sequences x, y of equal length n, and let

$$p_{\mathfrak{c},\mathfrak{d}} = \frac{\sharp \text{ pairs } (\mathfrak{c},\mathfrak{d}) \text{ aligned under } \pi}{n}.$$

Collecting these ratios in a vector $\vec{p}_\pi(x, y)$ for all pairs $(\mathfrak{c}, \mathfrak{d})$ we obtain the *empirical distribution of aligned letter pairs*.

The alignment score is a collapsed version of the information contained in the vector $\vec{p}_\pi$,

$$S(x, y) = n \sum_{(\mathfrak{a},\mathfrak{b})\in(\mathcal{A}^*)^2} p_{\mathfrak{a},\mathfrak{b}}\, s(\mathfrak{a}, \mathfrak{b}) = n\langle s, \vec{p}_\pi\rangle.$$

Example 5. For π given by

x	$\mathfrak{a}$	$\mathfrak{a}$	$\mathfrak{b}$		$\mathfrak{b}$
y	$\mathfrak{a}$		$\mathfrak{b}$	$\mathfrak{a}$	$\mathfrak{b}$,

we find

$$\begin{aligned}\vec{p}_\pi(x,y) &= (p_{\mathfrak{aa}}, p_{\mathfrak{ab}}, p_{\mathfrak{a}\mathfrak{G}}, p_{\mathfrak{ba}}, p_{\mathfrak{bb}}, p_{\mathfrak{b}\mathfrak{G}}, p_{\mathfrak{G}\mathfrak{a}}, p_{\mathfrak{G}\mathfrak{b}}) \\ &= (0.25, 0, 0.25, 0, 0.5, 0, 0.25, 0).\end{aligned}$$

Now let $X_1 X_2 \ldots$ and $Y_1 Y_2 \ldots$ be random sequences with i.i.d. random letters

$$X_i, Y_i : \Omega \to \mathcal{A}$$

with some given distribution, and let $\Pi_n^*(s)$ be an optimal alignment of $X_{[1,n]}$ and $Y_{[1,n]}$ under the scoring function s,

$$\Pi_n^*(s) = \arg\max_{\pi} S_\pi(X_{[1,n]}, Y_{[1,n]}).$$

Note that since the sequences $X_1 \ldots X_n$ and $Y_1 \ldots Y_n$ are random, $\Pi_n^*(s)$ is random, and furthermore, its choice is generally nonunique.

The associated empirical distribution $\vec{p}_{\Pi_n^*(s)}$ is thus a random vector

$$\vec{p}_{\Pi_n^*(s)} : \Omega \to \mathbb{R}_+^{|(\mathcal{A}^*)^2|}.$$

Theorem 6. *For almost all scoring functions s, the empirical distribution $\vec{p}_{\Pi_n^*(s)}$ converges almost surely to a unique point*

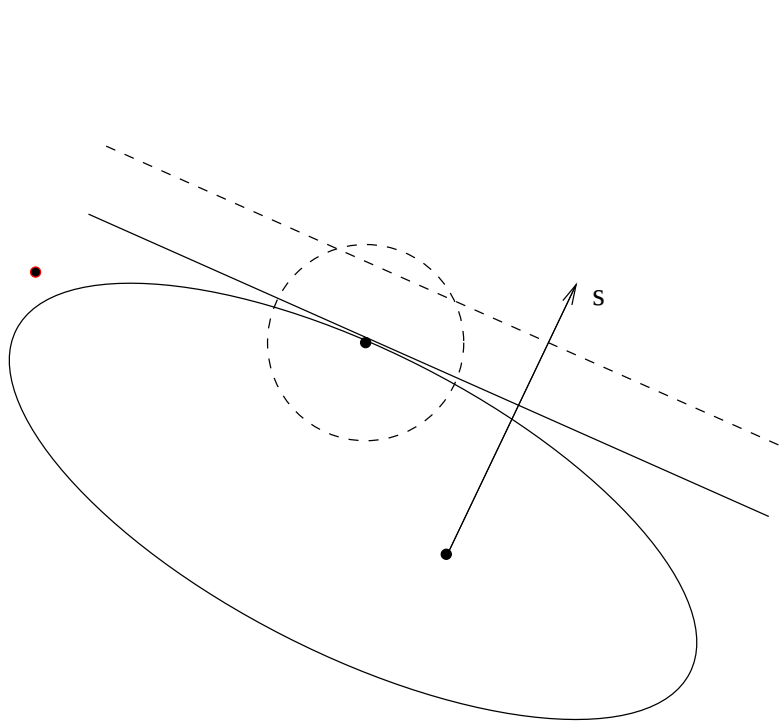
$$\vec{p}_s \in \mathbb{R}_+^{|(\mathcal{A}^*)^2|}.$$

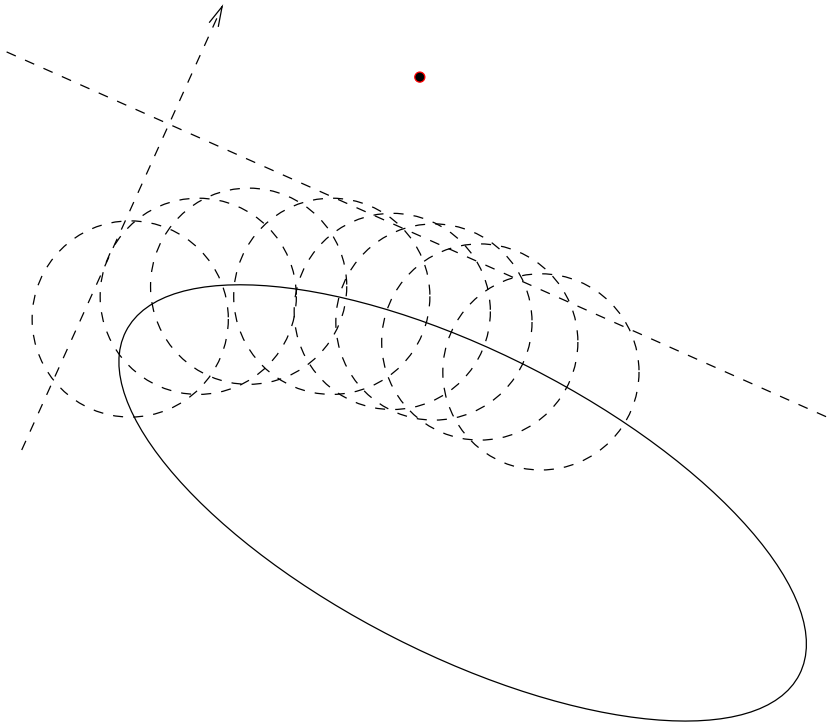
- In fact,

$$\mathsf{P}\left[\|\vec{p}_{\Pi_n^*(s)} - \vec{p}_s\| \leq \varepsilon \text{ for all choices of } \Pi_n^*(s)\right] \geq 1 - \mathrm{e}^{-K_\varepsilon n}.$$

- The asymptotic frequencies of the alignment microstructure are thus well defined.

- This allows for the design of more powerful statistical tests on the relatedness of sequences.


S



The empirical distribution may not converge when s is not chosen randomly!

- Take $\mathcal{A} = \{0, 1\}$, X_i, Y_j i.i.d. Ber(1/2) variables, and LCS scoring function.
- Subdivide the optimally aligned sequences into sections of length 3, e.g.,

$$\begin{array}{c|c|c|c|c|c} 1\ 0\ 0 & 0\ 1\ \text{G} & \text{G}\ 0\ 0 & 1\ 1\ 0 & 0\ 0\ 0 & \cdots \\ 1\ \text{G}\ \text{G} & 0\ 1\ 1 & 1\ \text{G}\ 0 & \text{G}\ \text{G}\ \text{G} & 0\ 0\ 0 & \end{array}$$

- One observes empirically that a positive proportion of triplets is of the form

$$\begin{array}{c} 0\ 1\ \text{G} \\ \text{G}\ 1\ 0 \end{array}, \begin{array}{c} \text{G}\ 0\ 1 \\ 1\ 0\ \text{G} \end{array}, \begin{array}{c} \text{G}\ 1\ 0 \\ 0\ 1\ \text{G} \end{array} \text{ or } \begin{array}{c} 1\ 0\ \text{G} \\ \text{G}\ 0\ 1 \end{array}.$$

- The first two correspond to the pattern $\begin{smallmatrix} 0 & 1 \end{smallmatrix}$ in X being aligned with the pattern $\begin{smallmatrix} 1 & 0 \end{smallmatrix}$ in Y, and the last two to the inverted situation.
- The first triplet can be exchanged for the second, and the third for the fourth without affecting $L_n(s)$.
- Interchange shifts weight from the pairing $(1, 1)$ to the pairing $(0, 0)$ in the empirical distribution.

Notes on the proof:

- Azuma-Hoeffding Inequality not applicable to empirical distribution $\vec{p}_{\Pi_n^*(s)}$, as changing a single letter in $X_1 \dots X_n$ or $Y_1 \dots Y_n$ may change the optimal alignment in a nonlocal fashion!

- However, Azuma-Hoeffiding is applicable to the function

$$(X_1 \dots X_n, Y_1 \dots Y_n) \mapsto \langle s, \vec{p}_{\Pi_n^*(s)} \rangle.$$

- The following is a compact convex set,

$$SET = \bigcap_s \left\{ x \in \mathbb{R}^{|\mathcal{A}^*|^2} : \langle s, x \rangle \leq \lambda(s) \right\}$$

- For almost all s (under the Lebesgue measure on $\mathsf{S}^{|\mathcal{A}^*|^2-1}$),

$$x^* = \arg\max_x \{\langle s, x\rangle : x \in SET\}$$

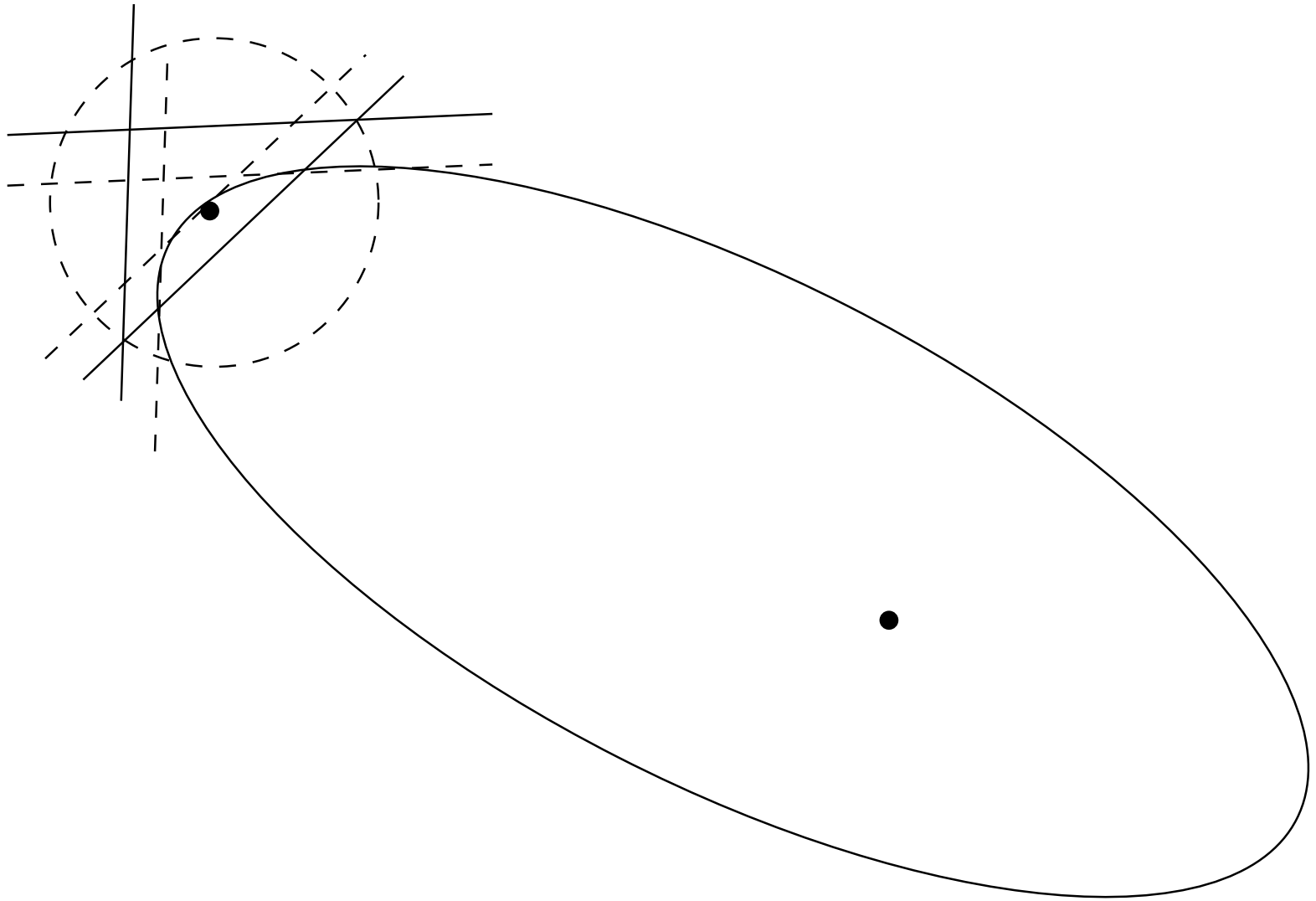
is unique.

- $\exists\ \eta < 0$, points $x_i \in \partial SET$, unit normal vectors $s_i \in \mathsf{N}_{x_i} SET$, and real numbers $\xi_i > 0$, $(i = 1, \ldots, k)$, such that

$$\{x : \langle s, x - x^*\rangle \geq \eta\} \cap \bigcap_{i=1}^{k} \{x : \langle s_i, x - x_i\rangle \leq \xi_i\} \subset B_\varepsilon(x^*).$$

- By Azuma-Hoeffing, all inequalities are satisfied with high probability.

- By Borel-Cantelli, almost sure convergence occurs.



Theorem 7. *Let s be such that $\vec{p}_s$ is unique (e.g., choose s randomly). If there exist $\mathfrak{a}, \mathfrak{b} \in \mathcal{A}$ such that*

$$\sum_{\mathfrak{c} \in \mathcal{A}^*} p_{\mathfrak{a},\mathfrak{c}} \left(s_{\mathfrak{b},\mathfrak{c}} - s_{\mathfrak{a},\mathfrak{c}} \right) > 0,$$

then

$$\mathrm{VAR}(L_n(s)) = \Theta(n).$$

- Close to proving order $\Theta(n)$ analytically in the generic case.

- Offers mechanism to verify the order $\Theta(n)$ by (non-brute-force) simulation.

More practical, weaker criterion

- Fix $\varepsilon > 0$, $\mathfrak{a}, \mathfrak{b} \in \mathcal{A}$.

- Given a scoring function $s : \mathcal{A}^* \times \mathcal{A}^* \to \mathbb{R}$, define symmetric difference

$$\begin{aligned} t_{\mathfrak{a},\mathfrak{c}} &= s_{\mathfrak{b},\mathfrak{c}} - s_{\mathfrak{a},\mathfrak{c}}, \quad \text{if } \mathfrak{c} \neq \mathfrak{a}, \\ t_{\mathfrak{c},\mathfrak{a}} &= s_{\mathfrak{b},\mathfrak{c}} - s_{\mathfrak{a},\mathfrak{c}}, \quad \text{if } \mathfrak{c} \neq \mathfrak{a}, \\ t_{\mathfrak{d},\mathfrak{c}} &= 0, \quad \text{if } \mathfrak{c}, \mathfrak{d} \neq \mathfrak{a}, \\ t_{\mathfrak{a},\mathfrak{a}} &= 2s_{\mathfrak{b},\mathfrak{a}} - 2s_{\mathfrak{a},\mathfrak{a}}. \end{aligned}$$

Theorem 8. *If* $\lambda(s) - \lambda(s - \varepsilon t) > 0$, *then* $\mathrm{VAR}(L_n(s)) = \Theta(n)$.

Simulation results:

$\mathcal{A}$	$\{0,1\}$	$\{A,T,C,G\}$
P[·]	$p_0 = 0.2, p_1 = 0.8$	$p_A = 0.4, p_T = 0.4, p_C = 0.1, p_G = 0.1$
s	LCS	BLASTZ
n	10^5	2×10^5
ε	0.5	0.9
$\frac{L_n(s) - L_n(s - \varepsilon t)}{n}$	0.0634	15.197
p-value	0.0102	2.4×10^{-4}

References

- R.A. Hauser and H. Matzinger. "Distribution of Aligned Letter Pairs in Optimal Alignments of Random Sequences". arXiv: 1211.5491.

- S. Amsalu, R. Hauser and H. Matzinger. "A Monte Carlo Approach to the Fluctuation Problem in Optimal Alignments of Random Strings. arXiv:1211.5489.

Thanks for listening!