A probabilistic framework for sequence distances & genome-wide (phylogenetic) pattern matching

Ali Osman Berk Şapcı[1], Siavash Mirarab[2,*]

{[1]: Bioinformatics and Systems Biology Graduate Program, [2]: Department of Electrical and Computer Engineering}, UC San Diego

Motivation

Estimating alignment-free distances between sequences of any length:

• Scalable • Accurate and sensitive (e.g., >15%) • Modeling uncertainty

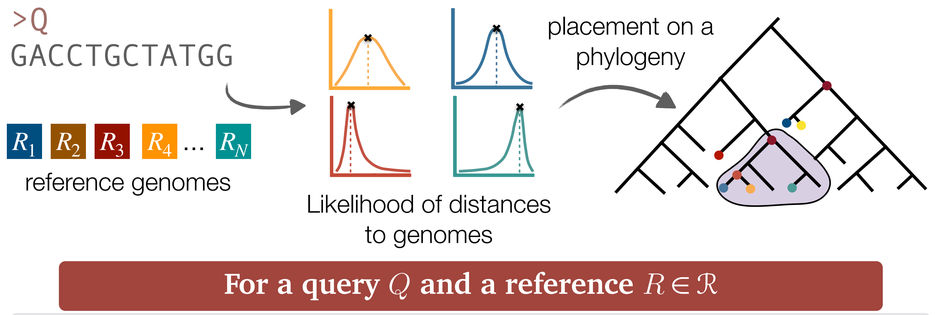
Applications of such a framework:

Metagenomics:
- Read-to-genome distances
- Phylogenetic placement

Comparative genomics:
- Horizontal gene transfer
- Ultra-conserved elements

Problem: alignment-free calculation of distances

Modeling the distance between. Q (e.g., read, contig, interval) and $\mathcal{R}$:



For a query Q and a reference $R \in \mathcal{R}$

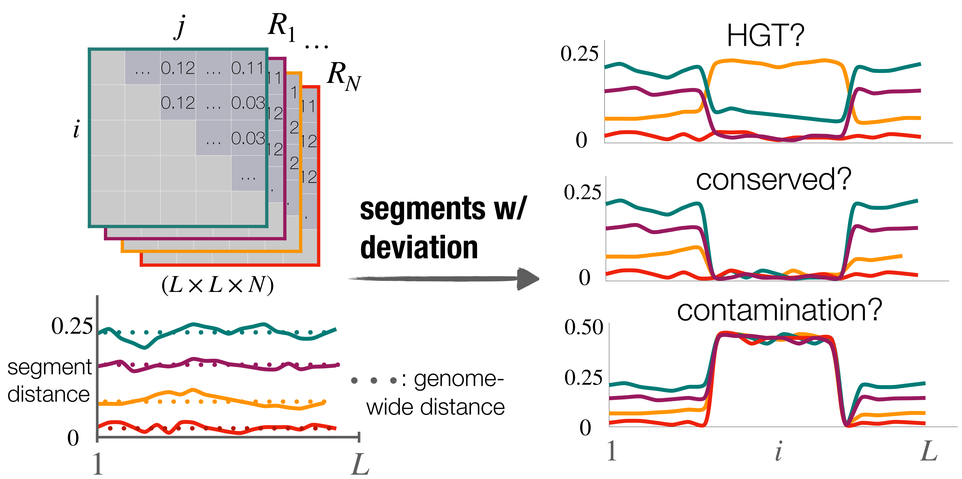
(i) Compute $\mathcal{L}(D_{ij}; Q, R)$ using k-mers, where D_{ij}=$\mathtt{d}(Q[i{:}j], R)$.

(ii) Estimate $D = \mathtt{d}(Q, R)$ and place Q on a phylogeny over $\mathcal{R}$.

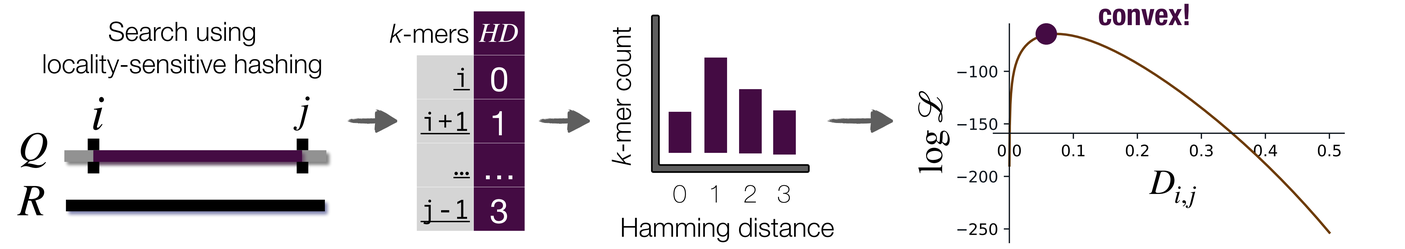
(iii) For all (i, j) and a threshold Δ, decide (in constant time) if $D_{ij} < \Delta$ (or $> \Delta$)

(iv) Given τ, find all "maximal intervals" with $D_{ij} < \Delta$ and $j - i \geq \tau$ $\mathtt{d}(Q[i' : j'], R) \geq \Delta$ for $i' \leq i \leq j \leq j'$, $(i, j) \neq (i', j')$

Detecting deviations of "local distances" from genome-wide distances:



Finding homologous k-mers and pseudo-likelihood formulation



Two key events:

For $x \in Q[i:j]$ **and** $y \in R$;

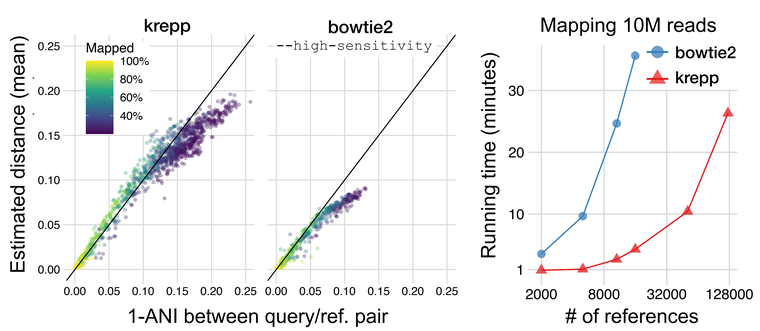
- They are d substitutions away:
$$P\big(\mathrm{HD}(x,y){=}d\big) = D^d(1{-}D)^{(k-d)}\binom{k}{d}$$
- LSH values collide (true positive rate):
$$P\big(\mathrm{LSH}(x){=}\mathrm{LSH}(y)\big) = \binom{k-h}{d}/\binom{k}{d}$$

misses: $u = (j-i) - \sum_{d=0}^{\delta} v_d$ matches: $\mathbf{v} = [v_0, v_1, \ldots, v_\delta]$

$$\mathscr{L}(D; k, h, \delta, u, \mathbf{v}) = P_{miss}(D; k, h, \delta)^u \prod_{d=0}^{\delta} P_{match}(D; d, k, h)^{v_d}$$

We define probability of observing u misses & v_d HD $= d$ matches based on **i)** subsampling rate **ii)** number of mutations **iii)** LSH match!

Estimating read-to-genome distances



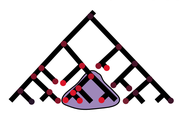
Phylogenetic placement

ancestral distances & likelihood-ratio test for placement:

$$\lambda_{LR} = \frac{\mathscr{L}_*(D; k, h, \delta, u_*, \mathbf{v}_*)}{\mathscr{L}_*(D^*; k, h, \delta, u_*, \mathbf{v}_*)}$$

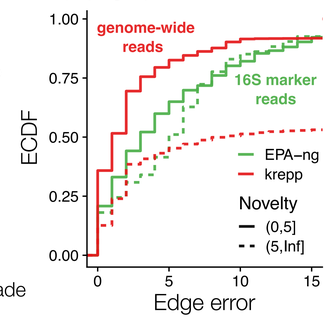
D: alternative distance

*: closest reference



place as sister to the largest indistinguishable clade

a phylogeny with 10,000 leaves



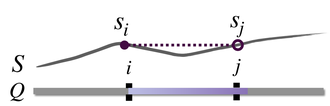
Distance-based segmentation & genome-wide pattern matching

MLE $\hat{D} = \Delta \rightarrow \ell'(\Delta) = 0$

If $\Delta < \hat{D} \rightarrow \ell'(\Delta) > 0$

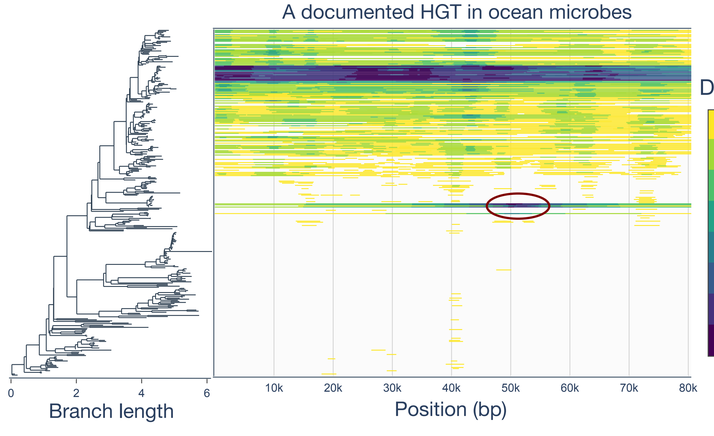
If $\Delta > \hat{D} \rightarrow \ell'(\Delta) < 0$

S: prefix-sum array for $\ell'(\Delta)$



- $\hat{D}_{ij} < \Delta$ iff $s_i > s_j$ **(constant time)**
- All *maximal* intervals **(linear time)**

A documented HGT in ocean microbes



krepp & gdiff

krepp for read-to-genome distances and placement:

GitHub: **Paper:**

gdiff for genome-wide local distances and segmentation:

GitHub: **Paper:**

Soon...