Motivation - Approximate Overlap Graph for Assembly

- Problem: Given a set of reads of roughly the same length, quickly identify pairs that are similar up to a certain number of shifts and substitutions (no in/dels)
- Motivation: New sequencing technologies (e.g. Illumina) produce
  - A large number of reads $n > 10^6$?
  - Reads of length $l > 100$ and increasing
  - Errors are substitutions
- Naive solution: Pairwise shift-align. Running time $O(n^2 l^2)$
- Geometric approximation algorithm finds all close pairs for n points in d dimensions in $O\left(n \log n + \left(\frac{1}{\epsilon}\right)^d n\right)$ (note the exponential dependence on d)

Discrete Fourier Transform

$$\mathcal{F}(\mathbf{x}) = \frac{1}{\sqrt{N}} \begin{pmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & \omega_N & \omega_N^2 & \omega_N^3 & \dots & \omega_N^{N-1} \\ 1 & \omega_N^2 & \omega_N^4 & \omega_N^6 & \dots & \omega_N^{2(N-1)} \\ 1 & \omega_N^3 & \omega_N^6 & \omega_N^9 & \dots & \omega_N^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_N^{N-1} & \omega_N^{2(N-1)} & \omega_N^{3(N-1)} & \dots & \omega_N^{(N-1)(N-1)} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{N-1} \end{pmatrix}$$

where $\omega_N = e^{\frac{2\pi i}{N}}$

- $\|\mathbf{x} - \mathbf{y}\|_2 = \|\mathcal{F}(\mathbf{x} - \mathbf{y})\|_2 = \|\mathcal{F}(\mathbf{x}) - \mathcal{F}(\mathbf{y})\|_2$
- $\mathcal{F}\left(\{x_{n-m}\}\right)_k = \mathcal{F}\left(\{x_n\}\right)_k \cdot e^{-\frac{2\pi i}{N} km}$

Random Projections

$$\begin{pmatrix} r_{11} & r_{12} & r_{13} & r_{14} & \dots & r_{1d} \\ r_{21} & r_{22} & r_{23} & r_{24} & \dots & r_{2d} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & r_{k4} & \dots & r_{kd} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}$$

where r_{ij} are i.i.d. random variables from e.g. $N(0, 1)$.

$$\|\mathbf{a} - \mathbf{a}'\|_2 = \frac{1}{k}\mathrm{E}\left(\|\mathcal{R}\mathbf{a} - \mathcal{R}\mathbf{a}'\|_2\right)$$

Discussion

- Disadvantages
  - No insertions/deletions
  - Shifts + substitutions should be $o(\sqrt{l})$
  - Floating point calculations
  - False positives (But no false negatives)
- Other thoughts
  - k-means clustering
  - Other geometric algorithms